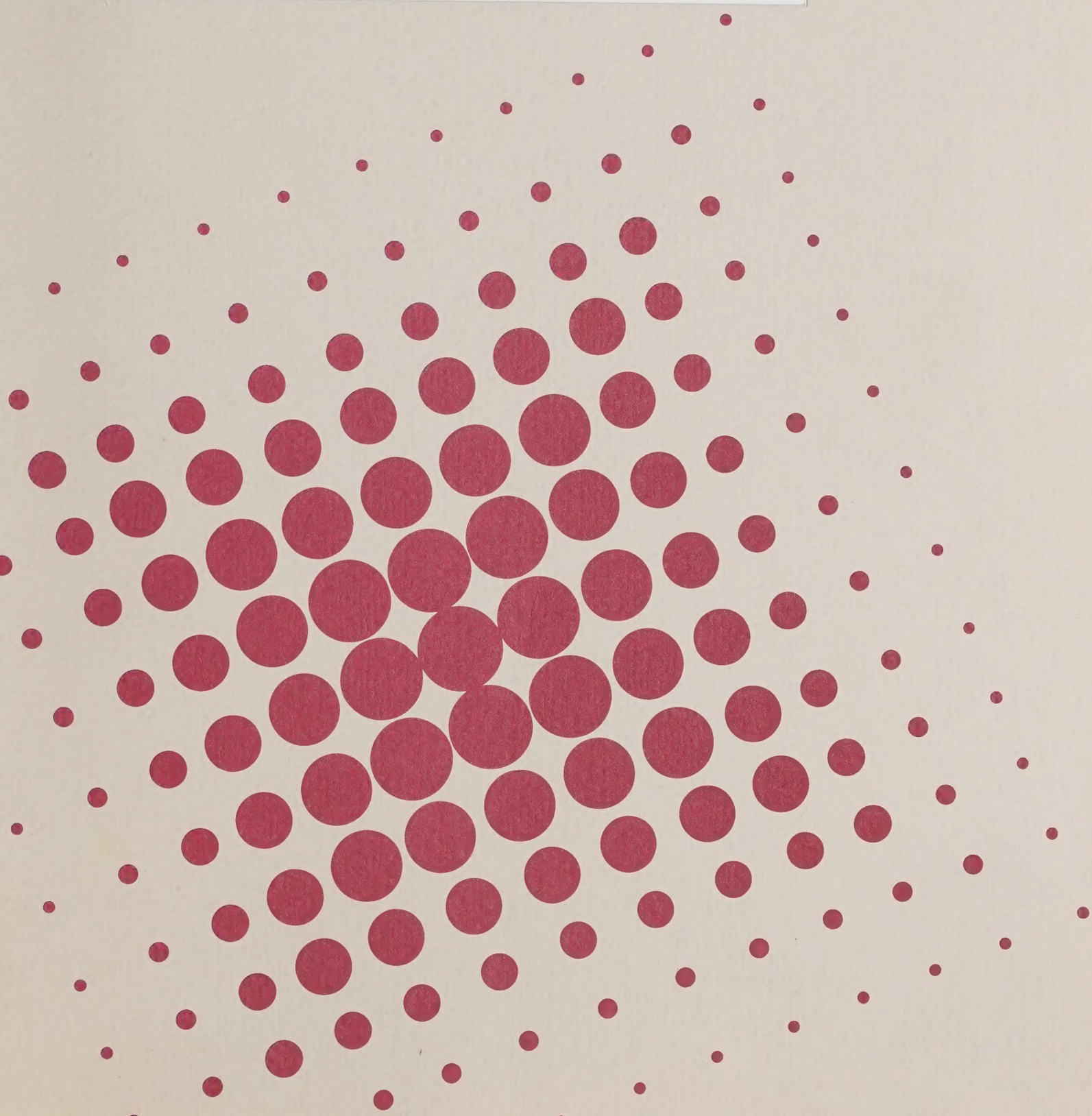
Brewster Kahle

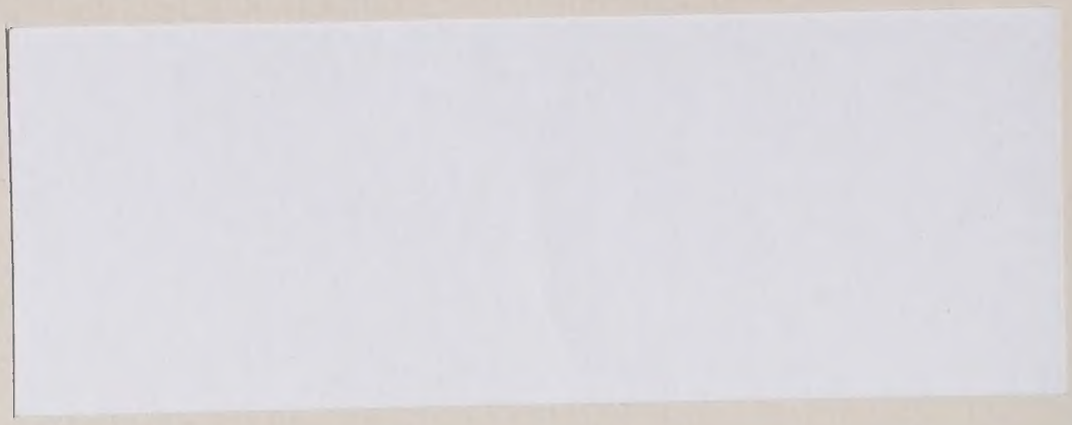
THE DIGITAL LIBRARY PROJECT

VOLUME 1: The World of Knowbots

(DRAFT)

Corporation
for
National
Research
Initiatives





THE DIGITAL LIBRARY PROJECT

VOLUME 1: The World of Knowbots

(DRAFT)

AN OPEN ARCHITECTURE FOR A DIGITAL LIBRARY SYSTEM

AND

A PLAN FOR ITS DEVELOPMENT

Robert E. Kahn and Vinton G. Cerf

Corporation for National Research Initiatives

March 1988

© Corp. for National Research Initiatives, 1988

Table of Contents

Summary i

1. Introduction 1

1.1 A Perspective 2
1.2 Technology and Infrastructure 11
1.3 The Digital Library Project 16
1.4 Spectrum of the Digital Library System 21
1.5 A Guide to the System 25
1.6 Applying the Digital Library System 29

2. The Architecture of a Digital Library System 30

2.1 Overview of Major Library System Components 34
2.2 Import/Export Servers 35
2.3 Registration Servers 38
2.4 Indexing, Cataloging and Referencing Servers 40
2.5 Database Servers 43
2.6 Accounting and Statistics Servers 44
2.7 Billing System 46
2.8 Representation Transformation Servers 47
2.9 Personal Library System 50

3. Knowbots and their Application 60

3.1 Overview 60
3.2 The Knowbot Operating Environment 63
3.3 Knowbots as Agents 65
3.4 The User Interface 67
3.5 Other Applications of the Digital Library System 71
3.6 Systems of Digital Library Systems 74

4. Implementation Plan 76

4.1 Phase one 76
4.2 Phase Two 83
4.3 Phase Three 85
4.4 Follow-on Plans 86

Summary

This volume describes an open architecture for an important new kind of national information infrastructure which we call the Digital Library System (DLS). The architectural framework includes the DLS functional components, the methodology by which the participating systems communicate with each other, and active, mobile software components, called Knowbots, which perform services for the users. Subsequent volumes will address detailed technical aspects of the architecture such as the design of Knowbots and the protocols required to bind the DLS components together. This research was carried out by the Corporation for National Research Initiatives to specify the overall structure and function of a DLS and to provide a basis for subsequent creation of an experimental system to evaluate the concept with real users.

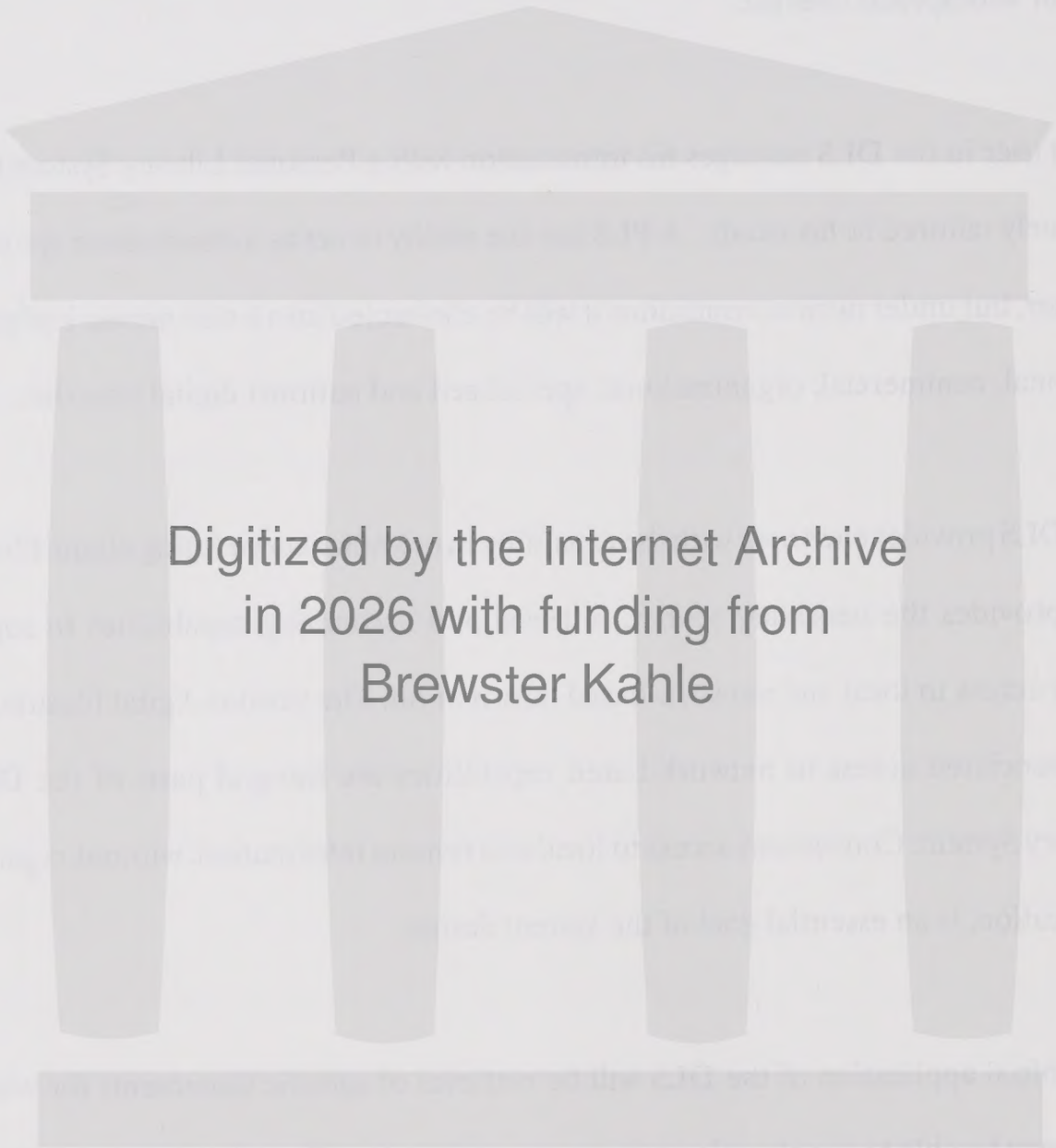
The term "library" conjures a variety of different images. For some, a library is a dim and dusty place filled with out-of-date texts of limited historical interest. For others, it is a rich collection of archival quality information which may include video and audio tapes, disks, printed books, magazines, periodicals, reports and newspapers. As used in this report, a library is intended to be an extension of this latter concept to include material of current and possibly only transient interest. Seen from this new perspective, the digital library is a seamless blend of the conventional archive of current or historically important information and knowledge, along with ephemeral material such as drafts, notes, memoranda and files of ongoing activity.

In its broadest sense, a DLS is made up of many Digital Libraries sharing common standards and methodologies. It involves many geographically distributed users and organizations, each of which has a digital library which contains information of both local and/or widespread interest.

Each user in the DLS manages his information with a Personal Library System (PLS) uniquely tailored to his needs. A PLS has the ability to act as a stand-alone system for its user, but under normal conditions it will be connected into a rich network of public, personal, commercial, organizational, specialized and national digital libraries.

The DLS provides each user with the capability to use other cooperating digital libraries and provides the necessary search, retrieval and accounting capabilities to support ready access to local and network-based information. The various digital libraries and the associated access to network based capabilities are integral parts of the Digital Library System. Convenient access to local and remote information, without regard for its location, is an essential goal of the system design.

The initial application of the DLS will be retrieval of specific documents for which a user may be able to supply only an imprecise description. For this purpose, we assume each retrieval request has a known target document which the user cannot describe or locate with precision, but can recognize when retrieved. Natural language and visual



Digitized by the Internet Archive
in 2026 with funding from
Brewster Kahle

https://archive.org/details/digitallibrarypr01unse_0

aids are used to assist the user in this process. However, the possible uses of a DLS are extensive and several innovative applications, discussed in the document, will be explored during the course of the project.

The potential utility for the Digital Library System technology is extremely high if agreement can be reached on appropriate standards and the relevant parties participate on a national scale. The efforts of multiple organizations such as computing equipment suppliers, publishers and other information providers and communications companies are needed to achieve the goals of the project. If successful, the results of this research offer the distinct possibility of enhancing productivity and should stimulate others to develop a vast array of new information products and services. The societal implications of success are very significant.

This document presents one practical path to the creation of such a capability. The benefits of this work depend on the outcome of the scientific research proposed herein. A plan is outlined for the development of an experimental digital library system which depends upon the active involvement of both the research community and the suppliers of equipment and services. The Corporation for National Research Initiatives looks forward to playing a leadership role in exploring the feasibility of this concept.

1. Introduction

This volume describes an open architecture for the development of a Digital Library System. The many users of such a system, even those with only limited or even no knowledge of information technology, can benefit enormously from quick and easy access to the information it contains. Its initial users will be drawn from the research community. However, the system is designed to accommodate a broad class of users (researchers and all others) in productive use of the digital library.

The Digital Library System design allows individual organizations to include their own material in the Digital Library System or to take advantage of network based information and services offered by others. It includes data that may be internal to a given organization and that which crosses organizational boundaries. This document presents a plan to develop such a system on an experimental basis with the cooperation of the research community. Finally, it addresses the application of a Digital Library System to meet a wide variety of user needs.

The productivity gains from having access to a Digital Library System are easily as large as those derived from internal combustion engines and electric motors in the early part of this century. Just as a car on an interstate highway is vastly more effective than one on a rutted dirt road, computer-based information "vehicles" can be made dramatically more effective given the proper operating environment. Computer and communica-

tions technology has made it possible for old fashioned, slow retrieval methods to be replaced by virtually instantaneous electronic retrieval. Each user of this technology can anticipate enormous potential benefit, but we lack the natural infrastructure to support this capability on a widespread basis today. This absence of infrastructure represents both a barrier and an opportunity of dramatic proportions.

1.1 A Perspective

Let us assume it is now 2003 and a little over a decade and a half has passed since the work which led to the development of a new national Information Infrastructure was started. The reality has surpassed early technological visions in unexpected ways although some of the more esoteric research ideas are still the subject of active investigation. To understand the profound nature of the revolution which has resulted from the establishment of this new infrastructure, we must reach back into history to trace the roots of human information processing.

In the early history of man, we had only an oral tradition with which to maintain our increasing fund of knowledge and recollection of history. Fathers handed down to sons the oral tradition of the tribe. The capture and rendition of this knowledge was a time-consuming process requiring frequent repetition to avoid its loss. With the invention of writing, we were able to speed up the process of information capture and simplify its reproduction. The price we paid to obtain maximum benefits for this improvement was a need to teach a larger community to read and write.

With the invention of paper and, especially with Gutenberg's invention of the printing press, the time and cost required to reproduce information was reduced dramatically. The invention of the typewriter brought personal printing within reach of a mass market, but the modification of printed documents to reflect changes and new ideas was still a laborious process which often required the re-typing and/or reprinting of the entire document.

Then dry-process reproduction methods were discovered and subsequently fashioned into copier products. This brought rapid reproduction of printed material to a mass market at an affordable cost. Once again, the time from the creation of a document to having multiple copies available for distribution dropped dramatically.

Computers brought yet another increment in flexibility and speed to the task of recording and sharing human knowledge. In the mid-to-late 1960's, time-sharing applications and networking, along with CRT displays, made it far less costly to prepare and alter documents before committing them to paper or other permanent media. Computer-supported text editing grew even more accessible and affordable with the emergence of microprocessors in word processors and then personal computers. By the early 1980's, most documents were prepared using word processing software running on personal computers.

An adjunct development, the computer-controlled laser printer, brought an additional level of convenience and flexibility to the recording of human knowledge. At very little cost, it was possible to produce fully formatted, multi-font documents with the same degree of revisability as one had with earlier, single-font systems. Products were developed which permitted the integration of imagery and graphics in digital form along with the textual components of documents. These products were called, collectively, "desktop publishing systems."

In the late 1960's and throughout the 1970's, the sharing of mainframe resources and information through networking was a popular method for distributing the cost of gathering and maintaining special information bases. Among the many such information services developed during that period, the bibliographic retrieval systems were among the most popular. Services such as the National Library of Medicine's MEDLINE, Lockheed's DIALOG and the Bibliographic Retrieval Service (BRS) became important reference tools for a variety of users. In the legal community, Mead Data's Lexis and Nexis databases became important resources supporting the preparation and evaluation of legal cases. These tools provide full text in addition to citations to their researchers.

The practitioners of Library Science, like many others dealing with increasing amounts of information, turned to computer-based methods for assistance. The Library of Congress, the Research Libraries Group at Stanford, the National Library of Medicine and

the Online Computer Library Center, Inc., joined together in a project to exchange information between their databases. This was called the Linked Systems Project.

Other information services, focused around the concept of remotely-accessible, on-line databases, emerged in this period. These included the Dow Jones/News Retrieval Service, the Data Resources, Inc., economic databases, the Thomas Register of companies and products, and hundreds of special databases reachable through Compuserve and The Source. These services, together, probably did not generate more than $300M/year by 1985; but there was a growing interest in access to information in this computer-processible form. Most of the information service providers were desired technology which would make more uniform the varied environment in which they had to work.

Networking, along with time-sharing and personal computing combined to form the technical support base for another important technology: electronic messaging. This technology emerged in the computer science research community in the 1970's and in the public domain in the 1980's. Electronic messaging further reduced the potential time delay for propagation of documents by allowing them to be sent electronically to the appropriate recipients. By the early 1990's, standards had been established and business relationships forged which permitted the interoperation of public and private electronic messaging services. The development and deployment of Integrated Services Digital Networking facilities, which emerged slowly in the marketplace, reached an average penetration of 30-35% by 1994. The bulk of this penetration was in the com-

mercial sector where over 70% of businesses had some form of electronic messaging service installed, while residential use had reached only about 15% of the market by that time. By the year 2000, this class of telecommunications had reached about 95% of the business market and about 35% of the residential market. Much of the early usage in the business market was attributable to "electronic data interchange" or "EDI" applications. In these applications, business documents (purchase orders, confirmation, shipping manifests and the like) were exchanged electronically along with electronic funds transfers.

The availability of a prototype Digital Library System (DLS) in 1992, with its innovative approach to intellectual property tracking, opened up a new publishing medium for information providers. In addition to the traditional book, magazine and newspaper markets and existing interactive database markets, the new Digital Library publications allowed the user to selectively view, reorganize and even update the contents for his or her personal use. Certain literary objects even had the ability to automatically update themselves with fresh information or to provide references to recent arrivals in the DLS.

Certain fundamental effects of the Digital Library System were the consequence of four distinct but strongly interacting developments during the 1990's. First, the conventions adopted for the representation of documentary material in the Digital Library were widely implemented by all vendors of combined document processing, database

and spread sheet software, by vendors of electronic desktop publishing systems and by public and private libraries around the country. The existence of such common conventions made it feasible for virtually any personal computer or workstation to access and use information produced by any other similarly equipped workstation.

Second, the continuing trend in reduced cost, increased power and memory in portable computers finally reached the point where real-time speech recognition applications became cost-effective. This meant that the transcription of voice to text became affordable to the business community in 1997 and to individual users by 2001. This same computing capacity, along with the development of high resolution, flat screen, touch sensitive displays, provided a basis for the recognition and transcription of hand written or pre-printed material as well. Direct interaction with a tablet/display and/or processing of scanned material became an affordable alternative to manual (i.e., keyboard) transcription.

The ready capture of imagery through high resolution scanning and telemetry added a third leg to the convenient creation, capture and processing of compound documents. Real-time manipulation and storage of material was also achieved.

Fourth, the incorporation of digitized sound, recorded or synthesized voice and high definition video sequences into documents stored in the Digital Library made it possible to combine most traditional forms of information publication into a common digi-

tal format. Conversion into and out of the digital forms and into the more traditional media provided bridges to older existing technologies. The structure and elementary content of printed material were determined during the scanning process.

The ability to register, store, catalog, search, retrieve and manipulate digital information in the library, combined with the variety of affordable media conversion capabilities available by the early 21st Century have led to a revolution in our social, economic and intellectual frameworks. Aided by computer-based Knowbots, easily reproduced and distributed computing cycles augmented human brainpower in the collection, use and creation of information in virtually every aspect of our lives.

Spurred, in part, by the focus of scientific attention on biology and biochemistry during the 1980's, and by the application of computer-intensive processing to non-invasive medical evaluations, the technology of the Digital Library System was applied to the capture and storage of high resolution magnetic resonance imagery (MRI), sonograms, X-ray and other similar diagnostic information. Increasingly detailed genetic analysis capabilities in combination with atomic level biochemical simulations have made it possible to carry out patient-specific bio- and chemo-therapy unthinkable in the past.

Massive amounts of detailed patient history, including the various kinds of digitized imagery, were stored in Digital Libraries around the country. This information provided a basis for epidemiological studies, simulation of experimental therapies,

analysis of the population for various health trends, tissue matching and statistical analyses for predictive or retrospective purposes. Coupled with the increasing use of computers for the fabrication of prosthetics, the conduct of surgery and the evaluation of drugs for therapeutic effects, Digital Library Systems are now playing a central role in health care in the 21st century.

Virtually all economic and social transactions are now recorded in Digital Libraries: property exchanges and documentation of ownership, the creation and dissolution of businesses and other legal entities, regulations, the judgments of courts and the acts of legislatures, births, deaths, marriages and divorces, the filing of intellectual property claims and the publication of intellectual works of all kinds are registered within the framework. Entertainment and advertising, product information and actual products, if representable in digital form, are lodged in and made available through these systems. Blueprints and designs for buildings, and other kinds of physical components are required to be deposited in Digital Libraries.

The exploration of information accumulated in Digital Libraries is now an essential part of our educational and research infrastructure. Computer-based tools for search and retrieval of information (including documents) are readily available to students at all levels. The results of manned and unmanned space exploration are indelibly recorded and made accessible as part of the system. Similar aggregations of information are accumulated daily from national and international high energy physics research

activities. Economic information, generated and captured in the natural course of daily transactions, is sorted, analyzed and mined by tireless Knowbots making their endless journeys through information space. Malthusian concerns about data overpopulation are easily solved by a combination of advances in high density storage systems and techniques which allow data to die a natural death.

The users of these systems draw upon natural language and visual capabilities embedded within them to find the information they need and to put it into a form suitable for further use. This information-rich, computer-aided environment has significantly changed our ability to organize into groups to achieve specific objectives. Our business organizations have taken on a much more fluid and "horizontal" character, now that the assembling and sharing of information has been made a natural side-effect of everyday interaction. New information-based products are introduced daily and are often discovered and used by other programs that serve our needs, without the need for our personal intervention at all.

Digital Libraries have now become such a pervasive part of everyday living that it's hard to remember what life was like without them. Like other infrastructure, one never really thinks about how it works, how it evolved or, how it is maintained, any more than one thinks about water, electricity, telephones and highways when they are readily available.

1.2 Technology and Infrastructure

Infrastructure plays a key role in the economic vitality of every nation. Viewed from an evolutionary perspective, infrastructure develops in response to the creation of new technologies. For example, the invention of the steam engine and its application in locomotives led to the development of railroads which, in the U.S., were instrumental in opening up the American West. Similarly, the harnessing of electricity and the invention of the light bulb preceded and motivated the development of a national power generation and distribution system. The invention of the automobile and the capability for its mass production ultimately led to the national interstate highway system which drove the evolution of suburban America. The telephone and the underlying communications technology led to a national telecommunications infrastructure.

Few inventions lead to the creation of infrastructure, but every so often, technology appears which drives this kind of development. Nearly every application which emerges at the heart of an infrastructure has an aspect of geographic dispersion and connectivity (e.g., telephone, television, roads, railroads, power generation). However, some technologies can form a kind of infrastructure without connectivity. Videocassette recorders are a prime example. Their penetration into the residential market is the basis for the cassette rental business which could not exist otherwise.

An important characteristic of infrastructure is simplicity of use. As with electricity, the user's view of the telephone, television, and automobile is essentially simple, although each of the underlying systems is quite complex.

Simple standards governing the use and application of an infrastructure also contribute to its utility in the social and economic structure. For instance, while power generation itself can be complex, ordinary 60 cycle, 120 VAC service is easily described and used to support an unending array of devices.

Computer technology, especially the personal computer or workstation, has all the characteristics consistent with infrastructure. PCs and workstations are widely dispersed in the geographic sense. Like many new technologies, their initial applications displace older methods of achieving similar objectives (word processing versus typewriters, for example) just as cars were thought of initially as simply horseless carriages. Once these displacement applications achieve sufficient penetration, though, it is possible to introduce quite new applications which have no previous counterparts. Moreover, the relatively recent development and spread of packet communication and internetting technology adds an important ingredient, namely connectivity.

These observations suggest that the ingredients are present for the creation of a new information infrastructure based on the wide and increasing penetration of computing and communication technology into the American social and economic fabric. Al-

though personal computers and workstations still suffer from user interface complexity, techniques have emerged (e.g., icons, mice, windows, natural language) which have the potential to simplify the use of computers considerably.

Infrastructure does not happen by accident. It is planned (either well or poorly) and deliberately created often with the direct involvement of the government. It is also preceded by a great deal of experimentation and research. The development of an information infrastructure will be no different in this regard. Unlike the industrial revolution which focused on the augmentation of human manual skills and abilities, computers offer the opportunity to enhance human cognitive capability and capacity. Over the last 40 years, the evolution of digital electronics, communication networks and computer-based applications has amply demonstrated the fertile potential of this technologies.

What is different at this juncture is that computers and digital storage technology are now readily accessible at reasonable cost to be applied to personal information tasks. At the home or office, and even on travel, the availability of computation is becoming pervasive. In the near future, shirt-pocket-size CD-ROMs (perhaps even writeable versions) will be commonplace. Use of networks to access remote databases will also continue to grow.

It is now likely that a substantial portion of the written information we encounter in the U.S. was, at one time, in computer manipulable form. Much of it has never been in that form, but the rate of production of information is so high that the more recent material significantly dominates that which has been produced in the past. Of course, the bulk of this information arrives as "marks on paper" in part because our information distribution methods are still dominated by the low cost and convenience of the printed medium.

Nowhere is the effect of this enormous influx of printed information more painfully recognized than in the research world where rapid access to relevant current and historically important results may make the critical difference between impasse and breakthrough. Finding relevant material, and even learning of its existence, is often a massive challenge. This problem is not unique to the research domain. It plagues virtually every information-dependent human endeavor.

Even if much of this information exists, however fleetingly, in computer processible form, it may not be saved or made accessible in that form. It is usually impossible for others to obtain access to it, even if they know about its existence. The need exists to establish an electronic information infrastructure to capture and allow this information to be made available (under the control of its proprietor). Computer and communication technology can be applied to augment our ability to search for, correlate, analyze and synthesize the available information. We describe such a strategy in this document.

Our initial results will make it possible to find and access copies of relevant documentation rapidly and in digital form, which will be a major improvement over current practice. Moreover, it will demonstrate an important example of information infrastructure which can provide the seed for a quantum leap in the way we handle all forms of information.

As information becomes increasingly accessible and fungible (in the sense that once in digital form, it can be readily processed by computer), the entire framework for compensating the creators of intellectual property may have to shift. At present, the basis for intellectual property protection in the U.S. is Patent and Copyright law. The large scale aggregations of information found on CD-ROMs and the selective access to information found in on-line databases may require substantial re-thinking of the ways in which the creators and owners of such information are compensated for its use.

There are many issues at stake in this area, not the least of which relate to the ease with which information can be replicated once in digital form and the rapidity with which large quantities of information can be processed (accessed, transferred, analyzed, integrated, etc.). Concepts of value and pricing and royalty for use of information could require considerable revision if the cost of such use is to remain within reason. One does not now pay an author a royalty each time a book is read. However, a royalty may be earned each time a song is played in public, though not in private. If a thousand books are combined on a single CD-ROM and the acquirer of the CD-ROM only in-

tends to read one of them, what sort of royalty arrangement is appropriate to compensate the copyright owners? How would compensation be extended for cases in which electronic copies are provided to users? In fact, the concept of copying or duplicating a work may no longer be the essential factor in calculating royalties since far more complex actions may now be taken on digital information.

These questions are not trivial in nature nor have many workable solutions been proposed thus far. It is critical that the interplay of various user and provider interests in information be considered and reflected in the design of the new information infrastructure.

1.3 The Digital Library Project

The digital library project is a broadly based effort to achieve coherent development of our national information resources. The existence of an open architecture for Digital Library Systems will provide the necessary structure for developing rapid access to existing information resources and for creating new information resources; some will be public, some commercial, some organizational and some personal. These will all be pieces of a larger composite library system if they adhere to the open architecture. Just as the highway system required judicious choices within each region and coordination at the boundaries, so will the Digital Library System. It can and should evolve to provide a seamless structure of access to information to encompass, in as far as practicable, the needs of all members of society.

By making it easier to use existing information resources, more people will utilize them naturally and hence the size of the user base will grow. The approach outlined here is to allow the user to stipulate simply what he or she wants to have happen and to let the system take the necessary actions. For example, to retrieve and print a specific document, the user would simply cite it by name. The library system would provide the necessary means for locating the information, retrieving it, and subsequently billing the user (the user could identify that he wants to know the cost before printing).

An overall architecture is needed to guide our use of such information in the future. The Digital Library System represents one practical path to the development of a coherent information base for the management and retrieval of data. The embodiment of this architecture and its assorted functions, protocols and standards in tangible experimental system will be a major contribution to the information infrastructure of the nation.

With the development of Digital Library Systems, enormous opportunities can be foreseen for creating and selling new products and services and for stimulating very significant increases in the demand for existing products such as workstations and print servers. One potential new product is a Personal Library System (PLS) which can manage all of a user's information needs. Personal and organizational data systems are logical extensions of today's myriad software packages and numerous services based on them can easily be envisioned. As with word processing and spread sheets, the use of a

PLS within the business community has the potential to streamline operations and improve productivity. For the research community, the ability to achieve quick and easy dissemination of results through electronic publishing will allow source knowledge to be propagated rapidly. For educational use, convenient and rapid access to reference material will quicken the educational pace and stimulate individual initiatives in teaching and learning.

To obtain the potential benefits of an information infrastructure, it is essential to promote the digitization of information and to insure that it becomes computer-accessible. Just as the widespread proliferation of the video cassette recorder has formed a technology base supporting an entirely new alternative to broadcasting, cablecasting and motion pictures, the provision of easy and affordable access to computer processible information leads to interesting new notions such as 1) "digital-back" publications as counterparts to hardcover and paperback books, 2) multi-media documents, whose elements may range over a substantial portion of computer-based publications, and 3) semi-automated retrieval services which can scan very large quantities of published and unpublished material for relevance to research and analysis.

Satisfying all the demands for access to on-line digital information is an overwhelming task for any one organization to undertake. Some of this information will be provided by existing suppliers and some will be created in computer-based form for the first time by new suppliers to the Digital Library. A significant portion of the material in the user's

personal library will be created by the user or collected from a variety of informal sources such as personal electronic mail, clippings and intra-organizational memoranda.

For the library to be a repository for personal and organizational information, it must have the participation of many different individuals and groups within each organization. By collectively engaging the creative energies of the many individuals and organizations in the country, a critical mass effort can be realized on a national scale. This broadly based participatory theme is an important aspect of the evolution of the Digital Library System.

The introduction of an information infrastructure is strongly affected by the environment from which it must emerge. There already exists an array of mass media types (newspapers, television, magazines, books) and some fragmentary electronic facilities such as electronic mail, and computer-based teleconferencing services, on-line financial, bibliographic, technical and business databases. Alternate technologies for mass publication of digital information are beginning to proliferate. For example, Compact Disk Read Only Memories (CD-ROMs) appear to be very attractive for many applications. These include the storage of large quantities of geographic, topographic and medical imagery (e.g. Defense Mapping Agency databases, NASA LANDSAT imagery, medical magnetic resonance imagery, etc.) and for large amounts of text and imagery as might be found in an encyclopedia, Patent and Trademark files, design

documents (architectural, aircraft, ships, integrated circuits, automobiles, etc.) or other reference volumes.

The development of an advanced information infrastructure must take into account a variety of existing and likely future interests and capabilities if it is to succeed. Publishers and authors must have reasonable incentives to make use of the new infrastructure. Existing libraries and their users must be able to make use of new technologies. Likewise, the educational system must be able to acquire and apply the products and services arising from a new information infrastructure if it is to serve their needs.

As viewed in this volume, the new electronic information infrastructure has a heavy computer-based aspect to it. Moreover, because the information is likely to be kept in digital form, the telecommunications industry (including telephone, television, local and wide-area networks, cable, fiber and satellite elements) will have an important role to play in support of access, retrieval and dissemination of digital information. For example, the planned development of Integrated Services Digital Networks and the longer term Broadband-ISDN could have a profound impact on the evolution of information infrastructure by providing easily used, variable rate, switched digital communication facilities. However, the role of the carriers could change in unforseen ways due to uncertainty in the regulatory arena.

1.4 Spectrum of the Digital Library System

A large amount of information is already available in computer-based form but is not easily accessible; therefore, relatively little use is made of it. Unless one already knows how to access such information, it may not be obvious even how to get started. Exploring databases for new information is at best a highly speculative process that is often expensive and unproductive. To the providers of database services, and the suppliers of user equipment, this situation translates directly into unrealized potential. Moreover, the vast majority of information that a user may ultimately wish to retrieve surely exceeds the currently available supply by a considerable amount. Without a system for convenient and widespread access to such information by unsophisticated as well as experienced users, it may never be economical to provide it. Until it is provided, however, widespread use may be stifled. Here we see a classic chicken-egg dilemma and hence progress on both fronts moves at a glacial pace.

The spectrum of possibilities for use of a Digital Library System system ranges from the tangible to the intangible, from the very specific to the vague and from the visual to the invisible. We depict one such range of possibilities by the series of six overlapping circles in Figure 1.

At the right-hand side of the spectrum, we denote fixed format documents intended to be read by people. These are generally assumed to be prepared for publication and have definite presentation formats. These documents are stored and retrieved in their

presentation form. They are guaranteed to be reproduced as they were originally created, subject only to scale and resolution limitations of the print server.

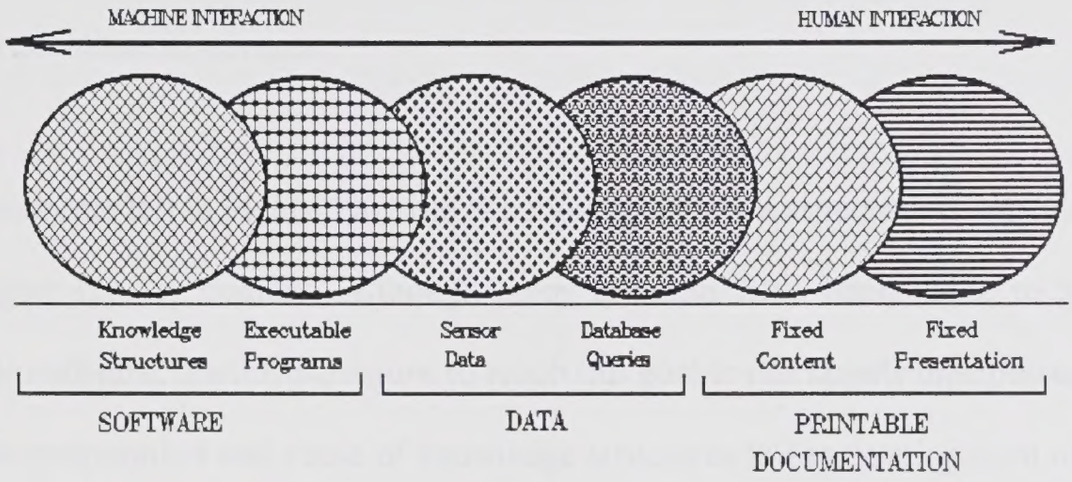


Figure 1 The Spectrum of Library Contents

Fixed Content, flexible format documents, shown just to the left of the fixed presentation documents in Figure 1, require the user or his system to specify how to present the information, assuming its content remains unchanged. For this class of documentation, the user might wish the text to be single spaced, double spaced, margins adjusted, page boundaries adjusted, fonts changed and so forth. This is in marked contrast with the fixed format documents, where no substantial visual changes of any kind are permitted.

In the middle of the figure are shown database queries and data of the kind collected from sensors. The system treats sensor data along with database entries as if they were new types of objects in the library; this treatment requires understanding the seman-

tics of objects in the library for the purpose of analysis and question/answering. When prestored answers are available without the need for searching documents, retrieval requests can be satisfied more quickly. Obviously, it will not be possible to anticipate all such questions in advance.

To the extreme left in Figure 1 are the two most speculative aspects of the spectrum of the Digital Library content. Although many attempts have been made to achieve reusable software, the infrastructure to reach this goal is still largely unexplored. Further, the preparation and reuse of knowledge structures in the development of intelligent systems is also virgin territory. This latter subject will be the focus of the second volume in this series.

The initial version of the Digital Library System will be tailored for the domain of printable documents (the two right hand circles in Figure 1). However, the underlying technology will be designed to allow evolution to cover the remaining portions of the spectrum. Ultimately, we see the library system encompassing the entire range of possibilities shown.

Even with this initial restriction on content, the span of possibilities for inclusion in the library is enormous. In the implementation plan (see Section 4), we discuss how the library system will be developed and how the supply of documentation can begin and expand.

Most users subscribe to a given information service to retrieve highly selective pieces of information. Rarely do they learn to use the full complement of capabilities available on that or any other system. Almost all existing on-line informational services support users that connect via simple alphanumeric terminals or PCs in terminal emulation mode. Most users are able to do little more than print a received text string or view it on a screen. The power of personal computers is rarely used to exploit further processing of received on-line information. With the exception of spread sheet programs that accept certain financial data obtained electronically, and mail systems that allow for forwarding, little or no user processing of received information typically occurs.

The underlying technology of the Digital Library System allows a user to access any available document within the entire Digital Library System. Using the PLS, he can modify a document in any way he chooses, incorporate it in another document, print it, search it or supply it as input to another program for further processing or display. Parts of the document can be extracted and manipulated.

Unlimited access to specific documents raises fundamental issues of intellectual property protection. A technological approach to this problem is outlined briefly in Section 3.1 of this report.

We plan to explore how to support vague and imprecise retrieval requests for specific printable documents while insuring that other well-defined requests are effectively handled as well. Requests for a specific manual, report, or equipment specification might be precise enough for the system to retrieve straightforwardly. The same might not be true if there was any uncertainty in the request. For example, if the author or title of a report were unknown, and only a general description of its subject was available, an intermediate process would be required to resolve the query further, to ask more questions of the user, or to produce a list of possible documents for selection.

1.5 A Guide to the System

A schematic description of the Digital Library System is shown in Figure 2. Its components are Personal Library Systems for the users, Organizational Library Systems for serving groups of individuals or activities, new as well as existing Databases stored locally and across the country, Database Servers to handle remote requests, and a variety of system functions to coordinate and manage the entry and retrieval of data. The system components are assumed to be linked by means of one or more interconnected computer networks.

Local requests for information, if not satisfiable by the local Personal Library, are dispatched to other, larger or possibly more specialized sources of information available through the network. A single inquiry may spawn tens to thousands of exchanges among

various parts of the full Digital Library System. This could easily happen if the system must first query several databases before responding to a particular inquiry.

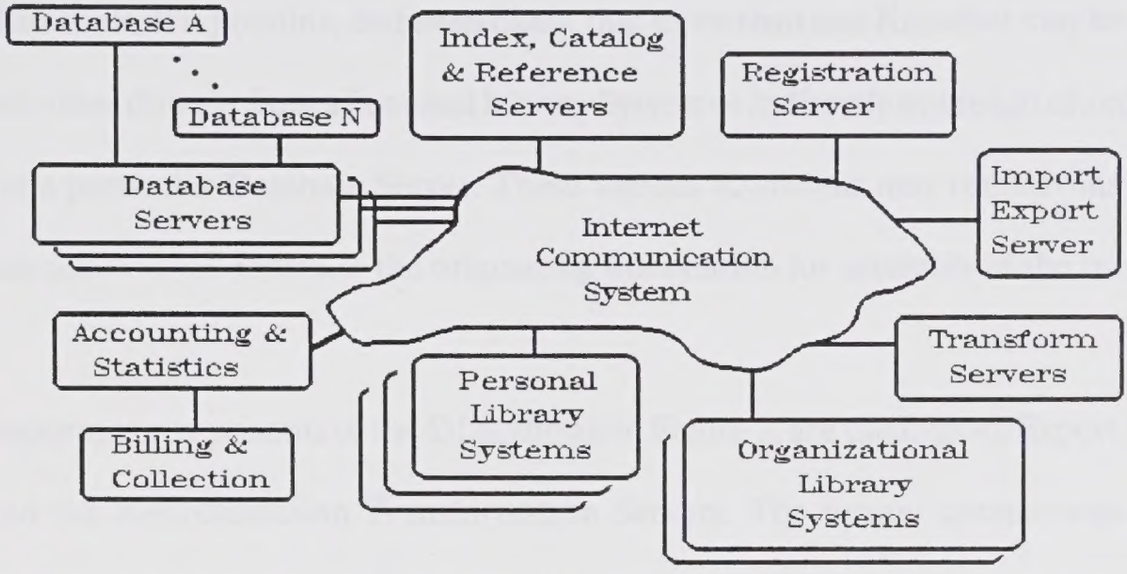


Figure 2 Structure of the Digital Library System

These exchanges are, for the most part, mediated by Knowbots, which are active intelligent programs capable of exchanging messages with each other and moving from one system to another in carrying out the wishes of the user. They may carry intermediate results, search plans and criteria, output format and organization requirements and other information relevant to the satisfaction of a user's query.

A Knowbot is typically constructed on behalf of a user at his Personal Library System and dispatched to a Database Server which interfaces the network to a particular database or set of databases. To accommodate existing database systems which are not capable of direct interaction with Knowbots, these servers can assist Knowbots in trans-

lating their information requests into terms which are compatible with the existing database's access methods. In the future, we expect to witness the development of databases systems with built-in mechanisms for housing and catering to resident or transient Knowbots. It is possible, and even likely, that more than one Knowbot may be dispatched either directly from a Personal Library System or indirectly as a result of actions taken at a particular Database Server. These various Knowbots may rendezvous at a common server or all return to the originating workstation for assembly of the results.

Two important components of the DLS, shown in Figure 2, are the Import/Export Servers and the Representation Transformation Servers. The former components are responsible for accepting new documents into the Digital Library System and for dispatching documents out of the system. The latter components convert documents from one internal representation to another. Depending on the nature of the output required, the obtained results may be passed through a Representation Transformation Server for conversion before being delivered. The results may be destined for either an originating PLS, a target PLS (or other workstation) designated in the original query or to an Import/Export Server if the destination is outside the particular Digital Library System "universe." For example, if the results are to be produced on CD-ROM and delivered physically to the user, this process will involve passage of the results out the Export Server.

When Knowbots and originating workstations or other intermediate information servers need assistance in finding information, they invoke Indexing, Cataloging and Referencing Servers by causing one or more Knowbots or messages from them to be dispatched there. The Indexing, Cataloging and Referencing Servers collectively contain information about the content and organization of the Digital Library System and help to identify which Database Servers should be accessed to respond to particular types of queries.

The Indexing, Cataloging and Referencing Servers are, in turn, kept up-to-date by the Registration Server which accepts new information into the Digital Library System. The Registration Server makes use of Indexing, Cataloging and Referencing Server(s) to determine where to store new information. The Registration Server also updates the Accounting Server so that providers of information can be identified and compensation for the use of information in the Library can be properly accounted for. Users of the Digital Library are also registered with the Registration Server and information about them passed to the Accounting Server, so that access to information and billing for its use can be supported.

Records of accesses and results are collected, by means of additional Knowbots, and reported to an accounting and statistics collection system for subsequent rating or analysis. The results of accounting collection are passed to a billing and collection system for further action.

1.6 Applying the Digital Library System

The Digital Library System will only be as effective as the various uses to which it is put. A few of these will be developed during the DLS project; the remainder will occur over time through the determination of motivated individuals and organizations. By way of comparison, it is noted that most of the applications for electricity came well after its introduction. However, a few key needs drove its early development such as its use for urban lighting. Applications using electrical motors came later on.

We can only begin to speculate on the many uses of a Digital Library System. However, several needs seem clearer than others; four of them are outlined in Section 3. One need is to examine and prioritize the contents of various publications which have been identified in advance and are known to be relevant to a given worker's field. A second need is to support computer-based design activities in which access to prior designs and their context or rationale is essential. A third need is to support research activities which involved searching for documents which contain relevant information and extracting critical portions for further and possibly detailed analysis. A fourth need is to link images and text for diagnosis. Many additional needs will undoubtedly occur to others.

2. The Architecture of a Digital Library System

Before describing specific features of the Digital Library System, it will be helpful to review some of the fundamental assumptions which strongly affect its design. Perhaps the most dominant of these assumptions are that the system is distributed, heterarchical, hierarchical, networked and strongly display-oriented. In addition, it must have an ability to interact with other autonomous Digital Library Systems that do not adhere to its internal standards and procedures.

The rationale behind the first assumption (distribution), in part, is that existing digital information sources are not physically collocated and that, as a practical matter, the Digital Library System design has to accommodate many geographically distributed components. The distributed system design does not rule out the centralization or at least concentration of resources where this meets pragmatic needs for minimizing operating costs, aggregating communications facilities, and so on. The important point is that the design forces neither centralization nor pure decentralization but accommodates both styles.

We assume that users will access the services of the Digital Library from powerful, geographically distributed and often locally networked workstations. This assumption places networking at the center of the distributed architecture. Even if all the data content of the Digital Library were centralized, its users cannot be.

Distinctions between entirely different (autonomous) library systems leads to at least one level of hierarchical structure in the architecture. Components which can interact among themselves using an internal set of conventions are distinguishable from the set of components which use an external conventions. Distributed, decentralized but hierarchically structured computer services seem to be a natural consequence of the organization of the present and foreseeable marketplace for the use of systems like the Digital Library. Computer services which cross the jurisdictional boundaries between organizations, or even between divisions or departments of one organization, require management structures for access control and accounting. Services which span multiple organizations typically exhibit two or more levels of hierarchical structure stemming from the necessity to draw boundaries around component operating and management responsibilities.

Another rationale behind the hierarchical structure of the system is to constrain the scope of the data management problems so that system growth does not lead to exponential amounts of database updating and consistency checking activity. Similar motivations often impose structure on otherwise unstructured telecommunication networks, for example.

The importance of scaling in all dimensions cannot be over- emphasized. The architecture must scale in sizes and numbers of databases, numbers of users, numbers of components, bandwidth of underlying data communication, varieties of archived content and variation in presentation media and access methods.

By deliberately treating parts of the Digital Library as distinct, networked components, it becomes possible to simplify implementations and to identify explicit protocol, management and control interfaces required to carry out the functions of the system. Such structuring also has the benefit of accommodating potential competition among multiple sources for the provision of products, services and functions which, in the long run, improves user choices and enhances the opportunities for growth of the Digital Library System.

The assumption that users will access and use the services of the Digital Library through powerful, display-oriented workstations is rooted partly in the observation that personal computing and graphics-based workstation technologies are rapidly converging. As costs drop, personal computer users tend to buy increasing capability at the same cost, rather than spending less to obtain previously available capabilities. Economics aside, another reason for assuming the use of high power workstations is the need to support multi-font text, graphics, imagery and, possibly, other modalities (sound and video, for example), if the full range of potential Digital Library services is to be supported.

Such reasoning does not rule out catering to "disadvantaged workstations," but these are treated explicitly with the realization that there is a potential loss of fidelity, functionality, or quality of service when accessing Digital Library services through these less capable devices.

The heterarchical assumption is motivated by the likelihood that more than one such Digital Library System will emerge as the national and global information infrastructure evolves. In the past, architectural designs for distributed systems often have been based on the assumption that there is only a single, monolithic, integrated architecture. Such assumptions usually lead to serious limitations on interactions between autonomous distributed systems and thus inhibit any ability for them to coordinate, cooperate and interoperate. Examples of such lack of vision may be found in many of the private and public electronic mail systems which make no provision for addressing messages outside the domain of the specific mail system in question. The Digital Library design specifically contemplates the existence of multiple instances of autonomously operating Digital Library Systems requiring compatible external interfaces. Each Personal Library System will also comprise multiple internal components which need to interact closely.

The second assumption motivating the heterarchical design stems from a belief that useful, self-contained, workstation-based, personal digital libraries are needed which can interoperate seamlessly with other internal or external library components of an

organizational, regional, state or national character. The system design supports cross-links among components at various levels in the structure and, in fact, makes heavy use of such linkages to achieve efficient interactions.

2.1 Overview of Major Library System Components

In the sections that follow, we will examine each of the major Library System components in turn, describing their functionality and relationships with other components. Figure 2 illustrates a top-level view of the Digital Library System. The rationale for the order in which these components are described is based on following a document [or, more generally, an object as it makes its way into the Digital Library and is then accessed and used.

The principal components of the system are:

i) Import/Export Server
ii) Registration Server
iii) Indexing, Cataloging and Reference Servers
iv) Database Servers
v) Accounting and Statistics Servers
vi) Billing System
vii) Representation Transformation Servers
viii) Personal Library System

In addition to these eight basic components, there are two fundamental concepts which are intrinsic to the interaction of these various subsystems. These concepts are Knowbots and Shared Icon Geography which are discussed in more detail in Section 3. The initial information in the Digital Library system is assumed to be material which was originally intended to be printed (including multi-font text, graphics, bitmapped im-

agery) or otherwise displayed in static form. In addition to books, reports and periodicals, the system can include other material such as electronic mail, VLSI designs and organization charts. However, the underlying concepts will be easily extendable to allow more ambitious kinds of information such as holographs and digital films. The initial formulation of the system is organized around printable information to give the project focus and a concrete development target.

2.2 Import/Export Servers

An Import/Export Server acts as a primary interface between the Digital Library System and the outside world. Contributions to and acquisitions for the Digital Library are presented through an Import Server. The method of interaction with an Import Server forms one of the most important interfaces in the system. An Import Server will be capable of accepting contributions to the Library in many forms. Contributions and submissions might arrive as part of an electronic mail message, as a CD-ROM, as a magnetic tape, as a PC diskette or even as a facsimile scan. The common denominator is that the information has been converted to some definable digital form. One of the most important steps in the Digital Library design will be the determination of how many and which arrival formats will be acceptable. Conversion from analog to digital form, while an important consideration, is outside the scope of the library project.

The arriving objects (e.g., documents) must come with additional information if they are to be successfully entered into the DLS. Among other things, the Digital Library needs to know the origin of the object (bona fides); the owner of it (especially if any in-

tellectual property rights are to be accounted for); terms and conditions for use, reproduction and access (including access control lists on an individual or organizational basis, for instance); descriptive information which might aid in retrieval; relation to existing information in the Library (e.g., part of a periodical series, book series, revision, etc.); and format definition.

Information which is not in a form which can be directly accepted at an Import Server will have to be prepared by services outside the Digital Library (an opportunity for any number of public agencies or private businesses). Similarly, Import Servers for particular classes of information might be implemented and operated or sold competitively.

An Import Server extracts the information relevant to registration from the arriving submission, packages it for processing by a Registration Server, and then forms and launches a Knowbot to deliver it there. At this point, the simple model is to send all of the information, including the actual submission, along with the registration Knowbot. This could prove impractical for significant contributions such as books. An alternative is for the registration Knowbot to carry only the information needed by a Registration Server and to carry references to the storage facilities at an Import Server for use when the information is to be transferred and incorporated into a database or catalogued by an Indexing, Cataloging and Reference Server.

An Import/Export Server also provides a basic mechanism for the equivalent of inter-library exchange services. It should be possible for several, otherwise distinct, Digital Library Systems to exchange information, queries, responses and library contents. Analogous to conventional inter-library loans, this capability is essential if the Digital Library System technology is to be independently proliferated to support a variety of products and services. Every effort must be made to assure that the architecture is free of the assumption that a single system is unique in the information universe. This does not rule out the need to tightly integrate some Digital Library components into a particular coherent system, but emphasizes the need to tolerate and accommodate diversity.

It is not yet clear whether the inter-library exchange facility can be implemented merely as an electronic message exchange or whether the interaction should also permit more immediate and direct forms of Knowbot exchange. The latter may require too much context sharing or accounting/billing and authentication mechanism to be implemented for essentially distinct Digital Library Systems. Additional research will be required on these matters. For the present, it is assumed that an electronic message exchange convention will be the basis for interactions among distinct Digital Library Systems.

All such systems, if they are to interact at all, must share a common name and address space to support message exchange. This could be provided by relying on international electronic messaging conventions which include provision of such a common name and address space for electronic mailboxes.

In addition to its import functions, an Import/Export server has the responsibility for exporting information (objects) from the local library environment to other environments. The latter may be other libraries or other presentation media (paper, CD-ROM, facsimile, etc.). An object may be exported either as the result of an action taken by a user (or a Knowbot acting on behalf of a user) or as a consequence of a request for service imported from another library system. Although the inter-library exchange mechanism is assumed to be based on electronic mail, other less general but perhaps more efficient choices are possible. Other media conversions (e.g., to print) may have to be handled in idiosyncratic ways.

2.3 Registration Servers

Registration Server(s) are responsible for 1) receiving messages from or hosting arriving Knowbots carrying new information (or references to new information) to be added to the Digital Library, and 2) registering new users, sources of information (databases) or other components newly added to the system.

One of the most important tasks of a Registration Server is to associate a unique identifier with any new object in the system. Ideally, it should never be necessary to re-use any identifier; thus the identifiers need to be allowed to increase in length. If identifiers are to be assigned by more than one Registration Server, methods must be invoked to assure uniqueness (e.g., by prefixing the object identifiers with Registration Server identifiers).

A Registration Server reports the existence of a new object to the relevant Digital Library component. If the object is a new user, this is reported to the Accounting System and to the Indexing, Cataloging and Reference Server(s) so that queries regarding that particular user can be properly answered. New information to be added to the Library is likewise reported to the Accounting system in the event that charges are to be associated with its access and use. A Registration Server may also supply a description of the charging algorithm to be used for this information. This might be as simple as a reference to a standard algorithm or as complex as a program for computing use charges for the particular item.

If it is readily apparent which database server(s) should house the arriving object, a Registration Server will so inform the Indexing, Cataloging and Referencing Server(s) and direct the Registration Knowbot to ferry the data to the appropriate Database Server. Alternatively, if the information did not come along with the registration Know-

bot, a Registration Server can form a new Knowbot to pick up the information from the Import Server and deliver it to the appropriate Database Servers.

Registration Servers interact directly with Indexing, Cataloging and Referencing Servers by providing them with an instance of the object being registered. An Indexing, Cataloging and Referencing Server determines which database can house the object (there may be more than one) and reports this information to the Registration Server. Other items, in addition to documents, which require registration in the reference database include, inter alia, all intra-library servers, users and other known Digital Library Systems.

2.4 Indexing, Cataloging and Referencing Servers

The principal function of the Indexing, Cataloging and Referencing Server(s) in the Digital Library System is to provide global cataloging and indexing services for the retrieval of Library content. The system is organized to support multiple, cooperating servers. It is also planned to accommodate alternative, specialized Indexing, Cataloging and Referencing Servers within this architecture to take advantage of new ideas and implementations without requiring the removal or replacement of existing services.

An important design issue will be the control of potentially open-ended interactions between Registration Servers and multiple Indexing, Cataloging and Referencing Servers to avoid network congestion and deal with the resulting multiple copy database update problem. Criteria for selecting among alternative Indexing, Cataloging and

Referencing Servers must be worked out, if several deal with the same or inter-related information. It is easier to deal with the case that knowledge about the content of the Digital Library is partitioned non-redundantly among several multiple servers. For instance, one server might specialize in cataloging and indexing electronic mail messages, another in books and a third in journals or other periodicals. Alternatively, if redundancy is to be supported, it might be based on multiple, complete, copies of identical indexing and cataloging information, rather than overlapping or partitioned components. Maintaining a consistent set of registration database copies is an interesting challenge in its own right.

Indexing, Cataloging and Referencing Servers are also used to locate services and users as well as information in the Digital Library. This function has an analog in the electronic mail domain in which name servers make it possible to find mailboxes associated with users. Search criteria for the name servers may be as simple as first and last personal names or complex conditional expressions, involving job title and/or function, company name, special interests (if known), locale and other identifying characteristics.

There are two distinct questions which can be answered by the Indexing, Cataloging and Referencing Server when it is dealing with Library content:

- "Here is the data, where should it be stored?"
- "Here is the kind of data I want, where is it?"

These two functions are, in fact, very similar and require the same, base level input information. Thus, any tools developed for one function can potentially carry over to the other.

Each Indexing, Cataloging and Referencing Server is capable of carrying out a repertoire of functions which can be invoked by Knowbots arriving at the Server. Knowbots arriving at a Indexing, Cataloging and Referencing Server will usually be performing one of several specific tasks:

- Cataloging/indexing of a new Library acquisition.
- Searching for a cataloged or indexed item.
- Collecting statistics about the content or usage of the Library.

When a new item is registered, a Knowbot is dispatched to a Indexing, Cataloging and Referencing Server for guidance in cataloging and indexing. The arriving Knowbot carries with it any key word or other cataloging and index terms that may have been assigned on publication (e.g., by the Library of Congress, the journal publisher, the author, etc.). It may also carry the actual item content so as to support cataloging and indexing algorithms which operate on the full "text" of the new item. Of course, the Knowbot also carries information such as the source (author), copyright owner (if any), International Standard Book Number (or other identification of this type), publisher, date, place (and time?) of publication. Both published and unpublished works could be included.

The indexing or cataloging information may vary depending on the nature of the new item. For example, arriving electronic mail would typically be indexed by origin, To: and CC: recipients, date and time of origin, unique message identifier, originating mail system, subject matter, and, depending on the Indexing, Cataloging and Referencing Server, by key words or user-provided search terms.

2.5 Database Servers

The design of the Digital Library System is intended to accommodate existing databases and database services and to provide a framework for new databases organized around the concept of Knowbotic information storage and retrieval. Database Servers bridge the gap between already existing, database services and the Digital Library System by providing support for resident and arriving Knowbots and exchange of inter-Knowbot messages. The principal tasks of the Database Servers are:

- To accept and store new information, and
- To house arriving Knowbots bearing queries

Some Database Servers may only provide the second of these functions as is likely to be the case if the actual database is managed and updated essentially outside the Digital Library System context. For database systems which are designed to operate within the Knowbotic paradigm of the Digital Library System, the functions of the Database Server may actually be combined with the database system itself. It is possible, of course, that these functions might still be supported by a separate Database Server for efficiency reasons.

Another motivation for including the Database Server in the architecture is to utilize new parallel processing technologies to speed the search and retrieval process for both new and existing database systems. Full text databases could be searched in their entirety at very high speed. Coupled with the Knowbot concept, such special purpose servers could revolutionize the utility of existing databases. To achieve this goal, it would probably be necessary to collocate the Database Server and the database system it serves so as to provide an economical but very high speed interconnection between the two. For existing databases, such a specialized Database Server would absorb the entire database so as to permit ultra-high speed and novel searching algorithms to be applied independent of its pre-existing computational base.

Such an intimate link between the Database Server and the database will doubtless require both technical and business arrangements, particularly in cases where the database is considered to be proprietary. Where such an arrangement proves infeasible, the alternative is to configure the Database Server so that it looks to the database as an ordinary user but provides all the of required framework for interfacing to the Knowbots of the Digital Library System.

2.6 Accounting and Statistics Servers

The function of the Accounting and Statistics Server is to collect and store data relating to the use of the Digital Library System and to send the accounting portion of it to the Billing Server. Information collected by the Accounting and Statistics Server in-

cludes not only retrieval data appropriate for billing purposes, but also statistics needed to guide operational decisions. Examples include information needed to identify capacity problems; profiles of information use (e.g., to identify the need to replicate data to reduce delay or increase transaction processing throughput); and inter-Knowbot message traffic (e.g., to determine when it would be more efficient for a Knowbot to be resident and exchange messages as opposed to moving Knowbots between a given pair of sites).

It is important to note that more than one Accounting and Statistics Server can be incorporated in the Digital Library both for redundancy and for load sharing. This means that any element of the Digital Library that produces data of interest to the Accounting and Statistics Server(s) must be configured to know to which server the data should be sent. To increase system integrity, the Accounting and Statistics Servers should be configured to accept data only from the appropriate sources and to raise alarms when data arrives from an unexpected source. Obviously, if redundancy is to be used to deal with various potential system failures, more than one Accounting and Statistics Server needs to be configured to accept data from a given source and these sources need to be configured to report to more than one Accounting and Statistics Server. This is a sensitive design area because the cost of sorting through multiple copies of accounting data collected at multiple sites is potentially very high.

The principal sources of accounting data are the Database Servers since they have direct access to querying Knowbots and their inter-Knowbot message traffic. This information is conveyed on a periodic basis (or based on the quantity of data accumulated) to the Accounting and Statistics Server. Other important sources of accounting data are the Import/Export Servers which process inter-library requests. In principle, the accounting for such queries should originate at the appropriate Database Server, but for inter-library reconciliation, the Import/Export Servers also capture traffic exchange information and pass this to the Accounting and Statistics Server.

The Registration Server is another source of accounting information since the registration of a new Library object or a new user often has accounting and pricing implications. In effect, most of the Servers in a Digital Library can be sources of accounting or statistics data, depending on the charging policy adopted by the operator of the system. An important area for agreement between two Library Systems will be their inter-library pricing and reconciliation practices.

2.7 Billing System

The Billing System generates invoices for use of the Digital Library System based on information it gets from the Accounting and Statistics Servers. The Billing System also needs to capture information about newly registered objects and users and may do this either through records sent to the Accounting and Statistics Servers by the Registration Servers or by direct exchange with the Registration Servers.

The details still need to be worked out, but it is possible that accounting data can be collected and delivered as objects like any others in the Digital Library. The mechanics of billing users and collecting revenues for service are still to be determined. By the time such a system becomes operational, direct electronic funds transfers may be the preferred collection strategy, but for the sake of backward compatibility, the system should also be capable of interfacing with a conventional lockbox service. This also implies that invoices may need to be sent either electronically (e.g., via Electronic Messaging Services) or on paper (via the postal service).

2.8 Representation Transformation Servers

The design of the Digital Library is posited on the assumption that only a few internal standard representations for library objects will be required. There will be a vast degree of heterogeneity in the actual sources of information to be placed in the Library and an equally heterogeneous collection of recipients with preferences as to the format of retrieved objects.

To avoid the need to build into the Database Servers the ability to accept or generate the entire panoply of possible object representations, the Digital Library employs Representation Transformation Servers which can accept a standard library object and convert it into any of several output representations for delivery to a user. Similarly, objects arriving at the Import/Export Server which are not in a standard library form may be converted at an appropriate Transformation Server.

It is anticipated that Transformation Services will be a lively area for competition among vendors of Digital Library products and services. Any number of such servers might operate within the context of a given Digital Library. Alternatively, the developers of such software might configure it to run in the context of a Personal Library System (see below) which would interact externally using standard object representations but could manage conversions internally using software acquired for this purpose or by means of exchanges with a Representation Transformation Server. Although the standard library representations have yet to be selected, a variety of potential representations into which or out of which it must be possible to transform already exist and will be used wherever possible.

The Association of American Publishers have adopted a version of the Standard Generalized Markup Language (SGML) as their preferred representation for the exchange of compound documents. Compound documents incorporate multi-font text and graphics in addition to raster or other bit-oriented images. If it is the case that most books and periodicals published in the U.S. will have an SGML form at some point in the process of preparation for publication, it seems reasonable that the Digital Library support this form as one of its internal standards.

In the international community, particularly in the International Standards Organization (ISO), a representation known as Office Document Architecture (ODA) is solidifying as an international standard. The National Science Foundation EXPRES

project has adopted a version of ODA as its preferred representation for compound documents. This choice is compatible with the X.400 electronic messaging format recommendation of the Consultative Committee on International Telephony and Telegraphy (CCITT). Indeed, X.400 can accommodate the transport of either SGML or ODA encoded objects.

A third representation of considerable and growing popularity in the U.S. is PostScript, developed by Adobe Systems, which comprises an executable language capable of very detailed descriptions of document presentation, page layout, imagery and fonts.

A fourth representation of potential interest derives from the American National Standards Institute (ANSI) X.12C committee which is working on standards for Electronic Data Interchange (EDI), focusing particularly on business documents such as purchase requests, purchase orders, bills of lading, invoices and the like. A related set of standards have been prepared by the ANSI X.9 committee for electronic funds transfer. This representation might be important to the Personal Library System if it is applied to tracking of personal or organizational financial transactions. Electronic funds transfer mechanisms might also be invoked within and between Digital Library Systems for the purpose of achieving royalty or other compensatory payments for access to and use of the content of these systems.

A fifth representation of increasing importance is facsimile (especially Group III and Group IV). A large number of documents are now received in that form and printed on thermal paper (or plain paper), but it is not far-fetched to capture this information in digital form for storage in the Digital Library. In the long-term, one can hope for better character recognition capability so that facsimile scanned documents can be reconverted to ASCII or some multi-font encoding.

There are, in addition, a large number of different word processing formats such as those used by Wordperfect, Wordstar and Microsoft Word, to name just three. There are also numerous proprietary document representations developed by industry. The Digital Library would rely on the Representation Transformation Server to deal with these various proprietary document encodings, translating them as needed into one of the several Digital Library standards.

2.9 Personal Library System

The Personal Library System (PLS) should satisfy two distinct needs in the architecture of the Digital Library System. The first is to provide a basis for a completely stand alone instance of a library system which can operate independently from the collection of other Digital Library Systems or even components of a given DLS. The second is to interact with the other distributed components of the DLS. Both of these requirements are treated in this section.

Figure 3 illustrates an abstract view of the internal structure of a Personal Library System. The horizontal layering shown is essentially notional. There is no attempt to portray with any precision, the vertical relationships among components.

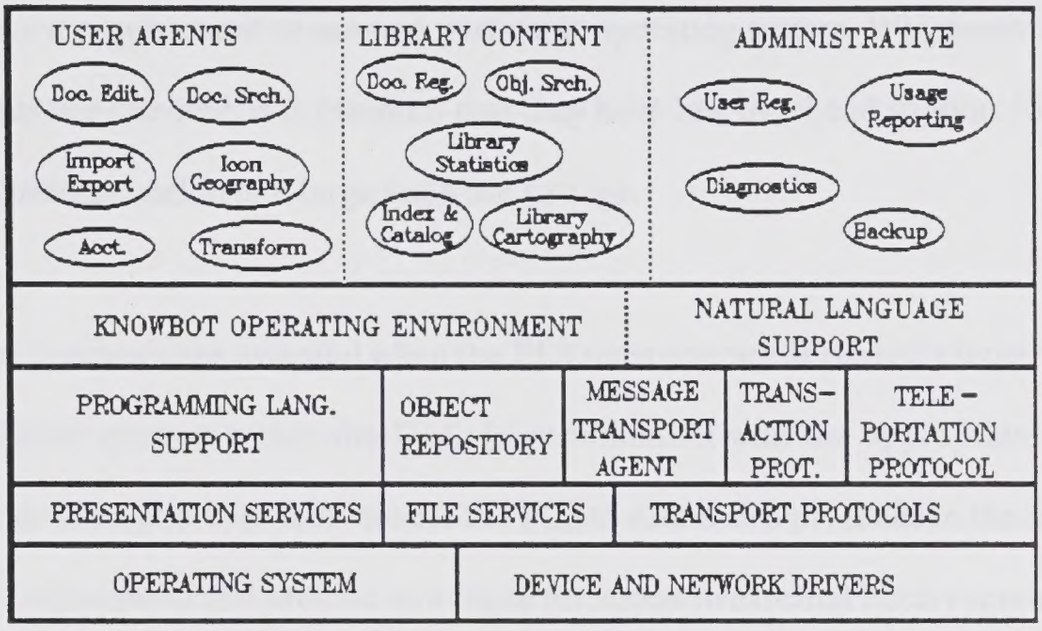


Figure 3 Personal Library System Structure

At the lowest level in the figure are the operating system and associated device and/or network drivers. It is not necessary for a PLS to be networked but it is increasingly common to find workstations interlinked on local area nets or at least capable of accessing dial-up telecommunication facilities. Although for convenience and simplicity they are not shown, included in the family of device drivers is support for common user interfacing devices such as keyboards, displays, mice and printers. These devices may even-

tually include audio input and output facilities and special high-resolution color displays to meet the "presentation" requirements of the contents of the library system.

The operating system will have to be capable of supporting multiple process execution. Many examples of such systems exist but the design of the Personal Library System does not impose a requirement to use one particular operating system. Whichever operating systems are selected, it is essential that they have low overhead support for interprocess communication and large scale file storage.

Transport protocols are essential when the PLS must operate as part of a larger collection of library systems within the DLS. In combination with the appropriate device drivers, the Transport Protocols enable the PLS to establish a presence in the rich networking environment and provide an avenue for access to external library services. Examples of the kinds of protocols which might be used include DOD TCP/IP/UDP, ISO TP/IP or other packet-oriented, multi-vendor protocols.

At least three application-related protocols are needed in the PLS if it is to occupy a useful place in a common networking environment. To handle electronic mail services, the PLS should support some kind of electronic message transport agent (MTA). This might be the DOD SMTP (Simple Mail Transfer Protocol), the Multimedia Messaging Protocol (MMP) or the CCITT X.400 (Mail Handling System) protocol. To support the exchange of Knowbots between the PLS and DLS components, a "teleportation

protocol" is needed. Finally, to support remote (or even local) message interaction between Knowbots, a "transaction protocol" is required. The inter-Knowbot messaging accomplished by means of the transaction protocol is distinct from the electronic mail interaction achieved using the MTA.

There is a two-fold need for the electronic mail capability in the Personal Library System. First, the PLS should be capable of assisting users in the searching, management and manipulation of their electronic mail. The PLS organization should attempt to accommodate this "under one roof;" however it is entirely reasonable for MTA functions to be provided by an electronic mail server external to the PLS but which the PLS can access to obtain copies of electronic mail intended for the PLS user. The second reason to have access to electronic mail is to provide an indirect, non-Knowbotic interface to external Digital Library Services. Distinct Digital Library Systems may not be able to share a common Knowbot Operating Environment but may want or need to exchange information. Electronic messaging technology offers one means for achieving this objective.

The Object Repository is a facility for storing the contents of the Personal Library System. The Object Repository is supported by the services of the Filing System (File Services in Fig. 3) which can be fairly conventional, but has its own organizational structure, access control mechanisms, indexing, storage and retrieval primitives. In the current design, all information stored in the Personal Library (and, in general, in the Digital

Library System) is object oriented. By this it is meant that the objects have "callable interfaces." Rather than knowing the details of internal representation of an object, it is enough to be able to call on the object to supply various pieces of information (e.g., provide a bit map representation for part of a document, provide information about the content of the object such as key words, provide information about the source of an object and so on). The representation returned from such calls does have to be standardized, to permit Knowbots to manipulate arbitrary objects and their contents. The motivation for this point of view is similar to the motive for the development of object oriented languages: simplified and standardized interactions with objects while allowing substantial variation in internal representations. For information in pre-existing databases, Database Servers are used to mediate and provide arriving Knowbots with an object view of the information. The concept of Knowbots is explored in more detail in Section 3.

At the present state of design, it appears that both Knowbots and the objects they deal with can be represented using object-oriented languages. Inter-Knowbot and Knowbot-object messaging is mediated through the Transaction Protocol. Although it is still not determined, the programming language support illustrated in Fig. 3 may turn out to be identical for Knowbots and objects.

There are a number of potential object-oriented languages which might serve for the representation of objects in the Digital Library System or for the representation of Knowbots. In the most general case, even Knowbots ought to be storable in the Digital Library as objects. Examples of existing languages include Smalltalk, Common Lisp, Common LOOPS and C++. The selection of a methodology for building Knowbots and even the determination whether an object-oriented language is essential are two of the highest priority research questions for the Digital Library Project to resolve.

A related representation concept called "hypermedia" or "hypertext" (a term coined by Ted Nelson) also needs to be taken into account. Originating with the early work of Engelbart on the On-Line System (NLS), the notion of threading text together in multiple ways with a variety of indexing and marking mechanisms has gained currency in the late 1980's. The notion has been picked up and expanded upon by others (e.g., Xerox with its Notecards experiment and by Apple with its Hypercard product for the Macintosh).

Ultimately, the Digital Library must implement methods for the creation, maintenance and extension of a rich collection of information registered in the system. Out of this will come facilities for easy browsing and association of related information. Whether and how notions such as hypermedia are reflected in the Knowbotic paradigm of the Digital Library is one of the intriguing research areas which will be exposed by the effort to construct and use an experimental system.

The Presentation Services subsystem concerns itself with the management of user interaction facilities and includes such functions as window management; icon, graphics and multi-font text rendering; linking of displayed constructs with screen coordinates to aid mouse utilization; and sound synthesis or capture. Most of these capabilities already exist and are assumed to be available for use in the Digital Library System.

Together, the Object Repository, Programming Language Support and Protocols subsystems provide the primary support for the Knowbot Operating Environment (KNOE) which is described in Section 3. The KNOE is a collection of software which mediates the creation, cloning, destruction, scheduling and migration of Knowbots. It provides an interface to the various underlying support services, including inter-Knowbot messaging, Knowbot teleportation and access to the Object Repository. Associated with the KNOE is a Natural Language Support subsystem which is built into the environment to make more efficient the processing of natural language by Knowbots. Natural language processing requirements arise from at least two sources: the content of objects in the Library and interactions with users.

Above the level of the KNOE and its associated natural language facilities, the PLS houses a variety of Knowbots whose functions can be roughly classified into three categories: user agents, library content and administrative. The Knowbots are illustrated at the top of Fig. 3, enclosed in ellipses. The ones shown are not intended to be exhaustive but rather to suggest the kinds of functions which would be present in a

stand-alone Personal Library System. Many, if not all, of these functions would also be needed for a PLS to operate in the even richer environment of multiple Digital Library Systems (or even one Digital Library System or even just another Personal Library System).

The document editing Knowbot interfaces with a user, making use of the variety of interaction support mechanisms discussed earlier. This Knowbot is capable of creating, interacting with and altering objects in the Object Repository, presenting them or otherwise rendering ("playing" in the case of audio output) their contents. The actual implementation of a compound document editor might involve a number of Knowbots, each with specific expertise in the manipulation of different classes of information.

The document searching Knowbot has knowledge of the contents of the PLS and is capable of interacting with the user to determine what information is desired. In the context of the larger Digital Library, the document searching Knowbot must have access to knowledge about the nature and whereabouts of non-local information. Such information, contained in the Object Repository, might range from precise identification of the location of a document to information only of other Knowbots to contact to assist with the search. A consequence of program or user interaction with the document searching Knowbot may be the creation of one or more new Knowbots which can assist in carrying out the search.

One of the more interesting concepts in the user interfacing part of the Digital Library is the notion of "shared icon geography." The idea is to extend the use of icon and window style interactions to linked three dimensional models of information space which can be shared across multiple PLSs. Distributed Library contents can be visually represented and can be organized in a familiar, physical, geographic or topographic fashion. Users might travel from place to place in this space, selecting objects for examination or organizing them in a new virtual space. Object representations might be linked or stored in various places in a fictitious information space. Search Knowbots, aided by Knowbots capable of producing three dimensional renderings, could organize information in accordance with user requests. Thus, the information landscape need not be uniform or constant for all users or even for the same user.

Accounting, import/export and transformation Knowbots would provide local services to the PLS similar to those contemplated in earlier description of principal Digital Library components with similar names. The accounting Knowbot, for instance, would keep track of the usage of or reference to personal library contents and, through the usage reporting Knowbot in the administrative category (see Fig. 3), identify usage for statistical or royalty reporting purposes.

The user agent Knowbots deal largely with users or on behalf of users and interact with Knowbots in the library content and administrative categories. The library content Knowbots assist in the registration of new objects (e.g., documents), deal with search-

ing the local object repository and capture statistics about the use or content of the library. The indexing and cataloging Knowbots are responsible for assisting in the search for or the installation of new objects in the library. As objects are added to the local library, the library cartography Knowbot keeps track of their presence. If the PLS is used to interact with other components of the Digital Library System, the cartography Knowbot captures data about the location and nature of these components and their content. Thus, the cartographic Knowbot can learn where to find objects or to find information about certain topics.

The diagnostics and backup Knowbots are tools for initiating special functional checks for proper system operation or for assuring that information stored in a Personal Library System can be reliably and redundantly archived.

In its Personal Library System mode, the user registration Knowbot is concerned with validating a local user for purposes of access control and possibly for accounting, especially in the case of access to information with associated usage fees. In the more general environment, the user registration Knowbot may be needed to validate incoming requests for information or to decide whether to host an arriving Knowbot. The Personal Library System is thus a microcosm of the larger scale Digital Library System.

3. Knowbots and their Application

3.1 Overview

A Knowbot is an active program capable of operating in its native software environment. Knowbots are present in each of the various components of a Digital Library System. They can be cloned, replicated, created, destroyed, can be resident at a given host system or can move from one host machine to another. Knowbots communicate with each other by means of messages.

Knowbots act as the primary medium of communication and interaction between various major components of the Digital Library System. They may even transport other Knowbots. Generally, a Knowbot may be viewed as a user Knowbot or as a system Knowbot depending on whether it directly serves an individual user or not.

A user Knowbot will accept retrieval instructions from a user and determine how best to meet the stated requirements, perhaps by interacting with other Knowbots and functional elements of the Digital Library System. Knowbots then proceed to acquire the desired information by accessing the appropriate parts of the library system. In carrying out this task, they may rely on intelligent indexing services provided by other Knowbots or perform actual text searching where needed.

One set of system Knowbots specifically attend to locally available library information. They take requests from user Knowbots and actually retrieve the documents from storage (or conversely store them away). Another set of system Knowbots attend to background and administrative tasks such as diagnostics, backup and accounting.

A class of trusted Knowbots called *couriers* have the special responsibility to look after selected objects on behalf of their authors or other owners of rights in the objects. A courier may be entrusted with responsibility for an entire database or a specific document or only a portion of it. Public domain documents which may be freely transmitted, used and copied will not generally require courier services. However, we view this as a special case of a courier which is passive. For purposes of this discussion, we assume that all documents are entrusted to couriers and never appear in the library system without an accompanying courier to protect the owner's rights. The combination of a courier and its accompanying entity (e.g., paragraph, document, database) is a controlled object in the system.

When a controlled object is provided to a user, all access to its contained entity is handled via its courier. If the owner of the entity originally wished to charge on a per use basis, the courier will be instructed to report such usage when it actually occurs, or to seek permission for use immediately beforehand and to deny access if it cannot be granted. Should a user wish to extract a portion of the controlled object, say for inclusion in another document, a new courier and controlled object would be created to

convey the information and to represent the owner's potential interest in the user's new work.

Certain Knowbots have a permanent status within each user's system and are known as resident Knowbots. Another class of Knowbots may be spawned dynamically for the purpose of carrying out a specific task and are deleted with the task is done. These are known as transient Knowbots.

Both resident and transient Knowbots have equal status within the Digital Library System while they exist. Should a resident Knowbot need to carry out a function at another site, it will cause a transient Knowbot to be cloned for that purpose. Transient Knowbots can also be used for system updating and for populating new user systems. In this case, they might be used as templates for creating permanent resident Knowbots at the destination and then deleted.

Although the details of Knowbot construction and operation are not fully determined, the structure of a Knowbot will be refined as we explore the design of the Digital Library System. Initially, however, we envision it to behave somewhat like a cross between a Smalltalk-like object and an expert system. Thus, we expect to use many of the attributes of object-oriented programming and rule-based systems initially. As experience with this type of active programming style develops, we would expect the Knowbot concept to evolve in both structure and capability.

3.2 The Knowbot Operating Environment

Knowbots are created, destroyed and otherwise managed by a Knowbot Operating Environment called a KNOE. The KNOE provides the context in which Knowbots function within a Digital Library System. It manages the system resources needed to support them and supports inter-Knowbot communication.

A cross section of the DLS is illustrated in Figure 4. It depicts the KNOE as an annular ring and the Knowbots as circles or spheres on its periphery. Each PLS is shown as a sector or wedge containing a portion of the KNOE and some Knowbots. The principal components of the DLS are also wedges in the figure.

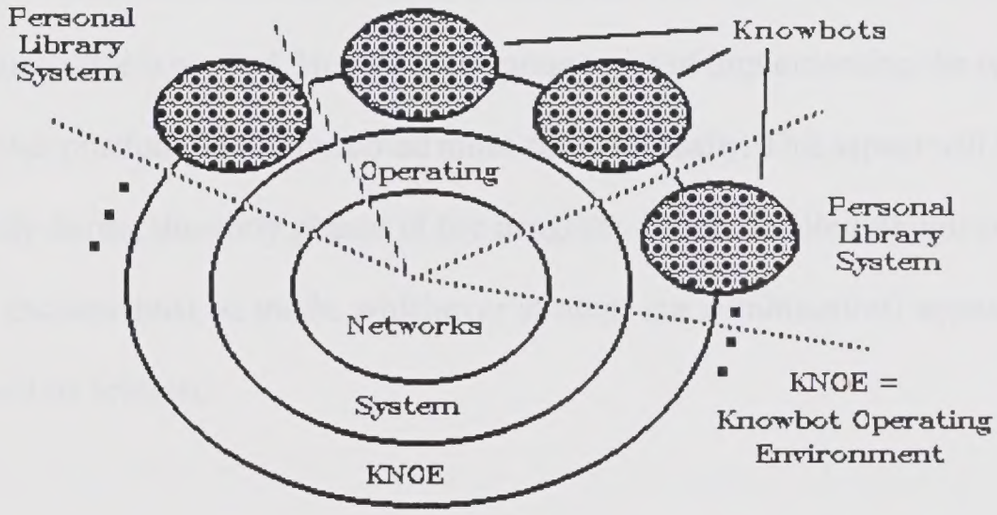


Figure 4 A Cross Section of the DLS

Each principal component of the Digital Library System contributes to and participates in the common Knowbot Operating Environment. Each local KNOE will know about all Knowbots in its local system and selected Knowbots elsewhere in the common KNOE.

Interactions between Knowbots are mediated by the KNOE. It assists in transporting messages between Knowbots in a given personal system and between systems. The KNOE will validate and authenticate messages when necessary. In a given local KNOE, any underlying message passing capabilities of the underlying operating system will be used by the KNOE in providing its layer of support.

Ideally, the KNOE could itself be created out of resident system Knowbots so that only a single architectural style is needed. However, the pragmatics of implementing the system may dictate that portions be programmed more conventionally. This aspect will be examined carefully during the early phases of the program. When detailed design and implementation choices must be made, whichever strategy (or combination) appears most desirable will be selected.

The resulting system will be designed for easy portability to other hardware and software bases. The ease of portability will depend on the extent to which the KNOE can be transported. If most of the KNOE is composed of Knowbots, then only a bootstrap version of the KNOE may be required. This is the minimum requirement on

the underlying hardware and operating system. If the entire KNOE is conventionally programmed, the demands made on the underlying hardware and software may be larger as well.

3.3 Knowbots as Agents

Knowbots may themselves be nested or defined recursively by drawing on the capabilities of other Knowbots, including themselves. For example, there might be a Knowbot created to handle compound documents. This, in turn might invoke separate Knowbots for handling text, images, graphics and even electronic mail (which might itself contain a form of compound document). It will be a design choice as to which Knowbots are visible to the user and which are hidden, in effect.

The top level of Knowbot in the system is called an *agent*. Initially, three agents are defined in the system. These are the user agent, the content agent and the administrative agent. Each agent consists of a set of resident Knowbots and each may request of the KNOE to generate transient Knowbots to assist with its work. At least initially, new agent types must be created outside the system.

The user agent consists of Knowbots for compound document generation and editing, document search and retrieval, document analysis, import/export, organizational structuring, accounting and authorization and interfacing with the user. Users deal directly with the user agent and each of its Knowbots, in turn, deals with the other agents.

The library content agent consists of Knowbots that handle document registration, indexing and cataloging, object storage and retrieval, storage management and icon geography, accounting and statistics. In addition, it contains a Knowbot to interface with other agents. The content agent is a system agent responsible for dealing directly with the object library. It receives input requests primarily from the user agent, but users do not communicate directly with it.

The administrative agent is concerned with tasks such as user registration, operations, diagnostics, backup and other similar functions such as financial analysis, billing and collection. Its main function is to support the other agents.

A typical user request for service might proceed as follows. The interface Knowbot would first determine the user's general intent and then attempt to capture what it believes is a valid user request. Let us assume the user wishes to retrieve a particular document but can only describe it generally in natural language.

The interface Knowbot would verify that the request was valid (this is largely research but simple tests can be used initially) and pass it along to the search and retrieval Knowbot to formulate a plan for satisfying the request. It might then invoke a strategy Knowbot, or a domain Knowbot to further refine its plan and then spawn one or more transient retrieval Knowbots. Each of these Knowbots might interact with other Knowbots to carry out its task.

If the requested document is not likely to be located in the user's personal database, but rather elsewhere in the system, the retrieval Knowbot is dispatched over the internet to other parts of the DLS. If the document is local, the Knowbot interacts with the storage and retrieval Knowbot in the database agent to hand off its specific task. Upon retrieval from the database, the document is supplied to the retrieval Knowbot representing the user agent after which it is ultimately made available to the user.

3.4 The User Interface

Knowbots have the primary responsibility for crafting the user's view of the library. The user conveys what he wants to see and how he would like the information presented. One or more Knowbots may then collaborate in creating the view.

The interface to the Digital Library System is essentially visual although we do not rule out other modalities such as sound. Knowbots as well as documents and controlled objects are depicted as icons and both may move dynamically in certain cases. Each Knowbot is represented by an iconic, three-dimensional symbol and its name, both of which may vary depending on the context.

The use of visual as well as logical recursion is intrinsic to the user interface. A Knowbot may be visible or invisible at the interface depending on its level of abstraction. For example, a simple search Knowbot which consists of a strategy Knowbot, an execution

Knowbot and a domain Knowbot may be represented to the user as a single virtual Knowbot that does searching or as some combination of these three.

Knowbots collaborate to depict distributed objects in the DLS. Messages are used to convey the necessary information from one Knowbot to another. Multiple users will also be able to jointly participate in a joint retrieval exercise and maintain consistent views no matter which user initiates or takes an action.

The object repository may reside at multiple locations, yet the user's view should enable a single coherent logical representation of the objects independent of their location. Two users collaborating in the library system should be able to share a combination of their views as a single coherent and integrated view of the system. We refer to this aspect of the system as shared icon geography.

One of the more important concepts in the Digital Library Systems is the idea of being able to share object representations, including the details of iconic presentation, with other parts of the system. For example, if a user has a Personal Library System which contains a number of objects, it should be possible to copy the iconic representation of these objects to another Personal Library so that two users can explore the same object space together. The resulting "shared icon geography," which includes both the details of iconic presentation and the cartographic relationship among objects, would permit groups of users to work concurrently in a common information environment,

coordinating the joint manipulation, examination and use of portions of the Digital Library's information space.

A particularly important issue is how to present retrieved information to the user when 1) the amount of it is inherently large or, equivalently, 2) when there are more than a few objects to be presented. This is fundamentally a research issue. We plan to seek a solution compatible with the use of shared icon geography. In addition, a simple way must be created to specify parts of the object space to browse. Electronic messages are a particularly good set of objects on which to start.

A specific request to "find the message I received 6-8 months ago about the design of the next generation workstation" may be too imprecise. Even if the system were told it was one or two pages in length with unknown sender, the Knowbot may still have to read each and every message to find the right one, if indeed it still exists (or ever did). If a few hundred messages should happen to fit the bill, the system might be unable to resolve the issue without first presenting choices to the user. The key question here is how to present this information most effectively.

As in the real world, the concept of "place" and "object" have meaning in the information space of a Digital Library. The iconic representations of places and objects in the Library are essentially multi-dimensional although they would be portrayed on a screen as two-dimensional projections of three-dimensional entities. Users interacting with

iconic representations of objects should be able to use familiar, real-world paradigms to manipulate the objects and maneuver in the places that populate the Digital Library.

Objects should be able to convey the notions of containment, emptiness and fullness. It should be possible to move objects, open and close them, enter them, move about inside of them, move other objects into them and so on. It should be possible to copy all or part of an object, assuming the user has the appropriate access rights. It should also be possible to designate portions of objects to be copied and transported elsewhere. This may be achieved by some combination of highlighting and annotating the appropriate portions of the object. Two very simple examples are 1) selecting bibligraphic information from a given text to be incorporated automatically in a personal data base and 2) collecting information typically found in address books such as name, address and telephone number.

The sorting and searching activities of Knowbot agents, working on behalf of one or more cooperating users, may result in the construction of three dimensional views of iconic objects found in the information space. Some of these representational ideas were originally explored by N. Negroponte of MIT in the Spatial Database Management System. Others are motivated by powerful notions of the visualization of active processes and the value of emulating common sense real-world behavior in the artificial information environment of the Digital Library System.

The realization of these ideas will require the application of leading technology in high resolution, color workstations, as well as research in powerful three-dimensional static and dynamic rendering methods, and techniques used in cinematography and television production to help "viewers" maintain context in complex visual scenes. To be effective, the user interface to the Digital Library will have to draw upon internal models of information space and options for navigating through it, techniques for animation, models of purposeful behavior and notions of goals and tasks to be accomplished. In short, the strengths of nearly every aspect of computer science, video-graphics, simulation and artificial intelligence must be marshalled to achieve the goals of the project.

3.5 Other Applications of the Digital Library System

Four possible applications of the DLS are described below. These are referred to as the Filter-Presenter, the Design Database Manager, the Researcher-Analyst and the Diagnostic Imager. Each of these uses would be implemented as an agent in the system.

The Filter-Presenter aids a user who is normally burdened by too much arriving information in the ordinary mailstream (e.g., magazines, newspapers, journals and electronic mail). If the material can be scanned electronically by the Knowbots, the user can be presented with only those aspects of the documentation he wishes to see. Of course, the user must first supply the Agent with sufficient guidance to carry out its task (including how he wants to see the results presented). Many research questions abound.

The filter may be too strong and therefore important items may be missed. Conversely, it may be too weak and the user will still be overloaded with irrelevant information. Irrelevant information may also be produced by a strong filter and relevant information missed by a weak filter, but both cases are much less likely.

The Design Database Manager couples a design program to an underlying database of relevant support information. It can also augment multiple design programs working collaboratively on a common design. In the case of VLSI design, for example, the elements in the database might be chip designs that could be used as subelements in a larger design. This could represent work underway by a team of designers. Or it could include standard designs such as simple microprocessors which may have limited use otherwise and could be used as pieces in a larger state-of-the-art chip. The advantage of this approach is that a new microprocessor design is not needed and users can be expected to be experienced in the use of existing designs for which software is already available. Alternately, the database manager could know about blueprints and how to use them and assist a user in retrieving and interpreting them.

The Researcher-Analyst assists a person who would normally search through large collections of documentation seeking specific types of information about a particular topic. It might identify hundreds of possibly relevant items that the user would have no time to explore. This agent could search all of them and develop information for the user depending on the nature of his research. For example, if the researcher was concerned

about the history of infrastructure, he might ask the system to locate as many documented examples of early uses of electricity as possible. This task might normally take weeks or months to accomplish manually but the agent using the library might accomplish it in minutes or less.

The Diagnostic Imager assists a person to find ways of binding textual or quantitative information with imagery. A reference to or selection of a given portion of an image or chart will automatically select the related textual information or vice-versa. This must be done in a well defined semantic context and not merely a geometric one.

Medical information needed to assist in patient diagnosis and evaluation covers an extreme range of modalities and levels of abstraction. From patient interviews to blood sample analyses to X-rays, CAT scans and electro-cardiograms, the Diagnostician is confronted with a rich and often perplexing array of information from which must be distilled an evaluation. In the context of the Digital Library System, the diagnostic task calls for access to a broad range of information which may range from specific information about the side effects of various drugs and chemicals to treatment protocols to indices of comparative medical imagery. The agent interacting with the user and accessing and manipulating the digital library content will rely on the use of Knowbots to transform symptomatic terms or analytic descriptions into appropriate keys for selecting useful imagery or to aid in searching for relevant treatments.

Finally, it should be noted that the structure of the DLS as a Knowbotic system makes it well suited for problems and applications that involve process control. A collection of Knowbots can be spawned to carry out a process control task and the library system architecture can be used to monitor its execution as if one were retrieving information from a more conventional library. The importance of this concept is noted, but it is not elaborated on further here.

3.6. Systems of Digital Library Systems

If an integral Digital Library System were to be constructed and placed in operation, experience predicts that evolution will result in other autonomous Digital Library Systems being generated in the future. The specifics of each system will surely differ from those of the others, and thus will arise the need for communication between these different Digital Library Systems. We call this inter-DLS communication.

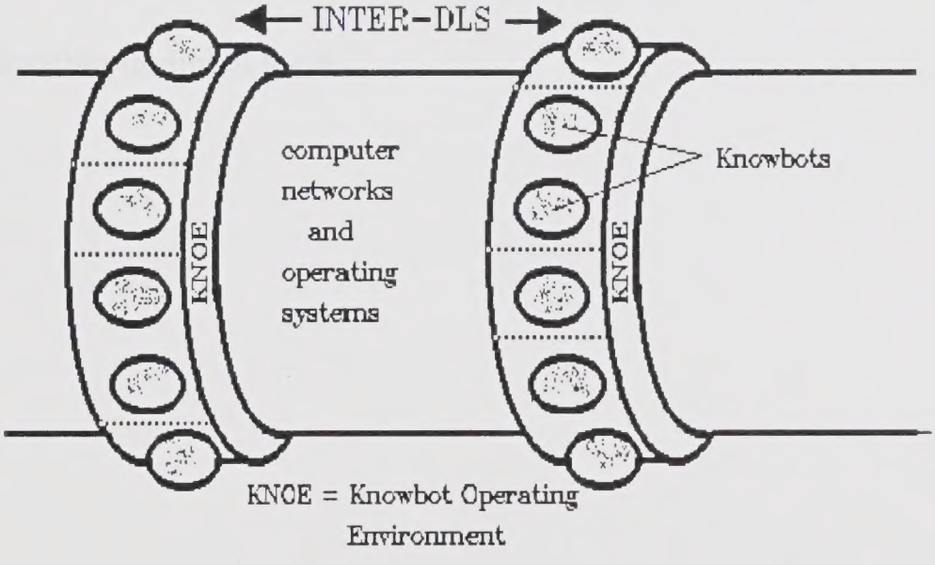


Figure 5 The KNOE Bracelet Model

We plan to define a language for inter-DLS communication which enables automous independent Digital Library Systems to interoperate. Most likely, this language will rely heavily on natural language as if it were supporting a normal user of either system. In this case, however, the user would be another DLS. The exact form of this language will be developed early in the project and will enable certain documents and requests to pass between systems. Inter-DLS communication is illustrated in Figure 5, where each DLS is shown as a separate bracelet on an internet-based, distributed substrate.

While a given DLS may be quite powerful in its own capabilities, the user should expect a more limited set of capabilities when multiple DLSs are involved. This may be dictated by administrative or legal restrictions dealing with enforcing copyrights as much as by technical limitations. For example, we assume Knowbots cannot be passed between autonomous DLSs and that certain documents may not be permitted to leave one DLS for another in digital form.

4. Implementation Plan

The architecture described in the previous section consists of eight significant functional component types interacting according to a set of protocols and a methodology which is observed by all the components. A three phase incremental development plan is presented here to achieve the library system program objectives. We first describe the elements which will be addressed in phase one which lasts an estimated 24 months, followed in succession by the second and third phases which last an estimated 24 and 12 months respectively. The activities in these various phases are described below.

4.1 Phase one

Four major tasks are presently envisioned for phase one. These are 1) the Knowbot Operating Environment, 2) the user interface, 3) populating the initial data bank for testing, and 4) Natural Language Text Searching. In addition, activities will be pursued in parallel to refine and further develop the overall system architecture, to explore one or more applications for the DLS and to consider matters relating to reasonable compensation for digital access and use of intellectual property.

4.1.1 The Knowbot Operating Environment (KNOE)

The description of the KNOE, the role of Knowbots, and their activities in the KNOE was presented in Section 3. This task will define the KNOE in detail, will develop a prototype system and implement three simple but functional Knowbots. Two separate subprojects are envisioned here to allow for competing views of the KNOE during this preliminary phase of the effort and to generate a total of six functional Knowbots.

The first of these two efforts, called KNOE-I, will concern itself with the issues of document creation and entry into the Digital Library System (DLS). It will focus on the user agent side of the library system and will implement Knowbots for the Document Editor, the Importer/Exporter and the Transformation Server. Using this system, documents prepared according to one standard may be manipulated by users with access to other standards or merged with documents prepared in other standards. The combined document will appear to the user as if it were a single properly merged document independent of whether the representation form of the document is changed or not. In general, the Transformation Server will convert the internal representation of a document without affecting the representation of the original copy. It will also insure that suitable notice is taken of the derivation of the work.

The Importer/Exporter will utilize a conventional database to demonstrate the issues involved with bringing new material into the Digital Library System or allowing it to migrate outside. It will also facilitate the incorporation of electronic mail or separate objects in the library system even if they were originally transmitted or received without any awareness of the Digital Library System.

The second KNOE effort, called KNOE-II, will concern itself with retrieval of documents in the library system and issues of management of the database itself. It will focus on the content of the library. In particular, it will implement Knowbots for search and retrieval, indexing and cataloging and a minimal registration function. At the time of initial deposit, each document will be cataloged and indexed for future retrieval.

The search and retrieval aspect will focus primarily on interactions with the database and with the user to a lesser extent. In both of these efforts, experiments with cooperating Knowbots and multiple KNOE's will be undertaken. Knowbots will move from one operating environment to another and will cause new Knowbots to appear at the other site by message passing. These activities will be under the overall supervision and control of the KNOE. Simple document handling scenarios will be executed.

4.1.2. The User Interface

This effort will implement a prototype visual user interface consisting of a shared icon geographic view of the contents of the library system and its components. Two components of the user interface will be the icon geography system which is responsible for interacting with the user and the library cartographic system which maps the relevant contents of the library and interacts with the object repository. These two systems interact with each other directly.

Initially, the functionality of these systems will be explored and demonstrated. Ultimately, each will be represented as separate Knowbots within the system, once the concept of Knowbots has been demonstrated.

4.1.3 Populating the Database

This effort will address those key aspects of the database having to do with personal, organizational (inter-organizational) and public information. Initially, the focus will be on populating an experimental database with objects that are publicly available or of organizational interest to be used for testing. After the Database Server is developed in phase two, network connections will be made available to existing databases.

We envision this effort will involve use of representation standards already in existence (or developed in the program, if necessary) and concentrate on collection and creation of an initial database. The equipment used will include scanners, optical character readers, and facsimile devices.

Public Documents - Our initial focus here will be to collect in digital form relevant standard documents from organizations such as ANSI, CCITT, ISO, IEEE, NBS or governmental agency pronouncements.

Organizational - Here we plan to focus on selected equipment manuals that are nominally available to customers or for internal use. This might cover hardware, software, and procedures for installation, use and repair. Eventually it might include brochures, pictures and specification sheets as well. Another organizational focus will be to collect and represent vita's of graduate researchers in the nation's colleges and universities. This will be carried out in cooperation with the university librarians who will each have responsibility for accurately representing their own school.

Personal - Finally, to make this effort interesting to the research community and to motivate some of their work, we plan to include selected research reports issued by the various universities which are not easily available otherwise. In addition, we shall include selected scientific and business publications based on their relevance and the willingness of the publishers to cooperate. Candidates are AAAI, IEEE, Scientific

American, and ACM on the scientific side, and Business Week, Harvard Business Review, Fortune and Forbes on the business side.

4.1.4 Natural Language Text Search

This effort will focus on demonstration of the use of natural language for search and retrieval. Initially, electronic mail will be chosen as a candidate to demonstrate retrieval based on imprecise English requests. When an object base with actual documentation is ready in the library, the domain will be expanded to include it as well.

This task will entail dealing with ungrammatical writing (but will not necessitate building a theory of such writing) and with understanding quite a bit about messages in general. Nonetheless, the domain appears to be relatively bounded.

The Natural Language System will be structured so that it draws upon existing state of the art technology but modifies it so that the system may be handed critical information it needs to do its job. Ultimately, this information will be provided directly by the Knowbots. In general, the Natural Language System will know about everyday English words. The system will receive a lexicon of relevant specialized words and their meanings, along with any additional helpful information. The system shall be structured such that new grammars and basic vocabularies may eventually be supplied to handle other languages.

In parallel with the above activities we expect several organizations to begin work on the development of a Personal Library System which provides user access to the library. These efforts may be simply to interact with the external development efforts, provide one or more individuals to work on them or begin an internal effort.

Throughout phase one we shall need to maintain, refine and update the architectural specification of the library system. At critical junctures, preliminary protocol specification documents will be produced for the critical interfaces between the principal components that were shown in Figure 2, for representation standards as well as for normal system capabilities such as Knowbots. This will be a living document which shall serve as the "sheet music" for the entire effort.

During this phase we will explore several possible applications of the Digital Library System along the lines indicated in Section 3.3.6. Initial designs for one or more of these efforts will begin. We shall also explore concepts for handling the reasonable compensation of intellectual property owners, since fragments of their as well as entire works may be involved.

4.2 Phase Two

In the second phase, we plan to begin development of a Personal Library System based on the research efforts in the first phase. In addition, we expect several industrial organizations to actively participate in the process so that a commercial source of the technology will ultimately be available.

The initial system will be based on a powerful workstation with local disk, scanner, printer and high resolution display (plus mouse and of course keyboard). Eventually, it will be graphics capable and be equipped with facsimile, low-cost personal high density storage such as a CD-ROM, plus an acoustic subsystem (including speech) and video.

Research will continue on several fronts. First, the Knowbot research will be expanded to include exchanges between Knowbots and making Knowbots work in collaboration with the object base. This effort will include Knowbots with domain expertise and the ability to do simple domain related problem solving.

The registration Knowbot will be fully developed and outfitted to handle users and services as well as documents. In addition, it will be expanded to be knowledgeable about an organizational level as well as a personal level of objects. It will also have a mechanism for internal collection of accounting information.

Research will also be undertaken to develop a Knowbot which understands organizational structure and documents associated with it. The structure of an organizational library system will be developed and simple exchanges between organizational and personal Knowbots explored. The Personal Library System technology will provide the basis for the organizational system as well; only the contents of the two systems will differ.

The task of database population will continue as a larger and richer set of documents is added to the object base. We expect this task to focus primarily on the more sophisticated elements of the existing documents such as equations, graphics and images. However, additional documentation will be added as appropriate for the research to be conducted.

The natural language capability for text search will be improved and a user front end will be incorporated which relies on a common body of natural language software. Experimentation with sample retrieval requests will take place interactively.

A Multi-Processor Database Server will be developed to access remote Database Systems. Initial experiments will be conducted with one or more cooperative suppliers of information. Candidates are the National Library of Medicine and Dow Jones. The Database Server will incorporate much of the Personal Library System software but be outfitted to operate several orders of magnitude faster than the personal workstation

and support multiple users. The server will draw directly upon the results of the Personal Library System research and apply it to multi-processors.

Finally, an effort will be undertaken to focus on some of the practical and administrative considerations such as tools for billing and collection, diagnostics, back up, etc. Also, those aspects which allow new capabilities to be added on the system to be reconfigured over time will be included here.

Sometime during Phase Two, we expect to obtain a working prototype of the initial Digital Library System and to begin making it available to selected members in the research community for experimentation and as an object of research itself. In addition, the development of one or more of the applications will be undertaken.

4.3 Phase Three

The components of the Digital Library System (DLS) will be fully integrated and a quasi-operational DLS will be created during this phase for research purposes. The existence of the system will serve to expand the Digital Library System to more users and a larger set of documentation. We expect to gain operational experience and to incorporate several additional public and private systems such as NTIS, Westlaw, Lexis/Nexis, Compuserve or Dialog.

An organizational prototype will be created working with one or more groups and experiments with inter-organization exchanges explored. Examples of inter-organizational exchange requirements will be developed and simple interactions carried out. These might include access to manuals, student vita, electronic mail, or open memoranda. Acoustic Input and Output including speech and other audible sounds will be incorporated into the user interface and further work on the shared icon geography carried out for dealing with complex representations. One or more applications will be substantially completed.

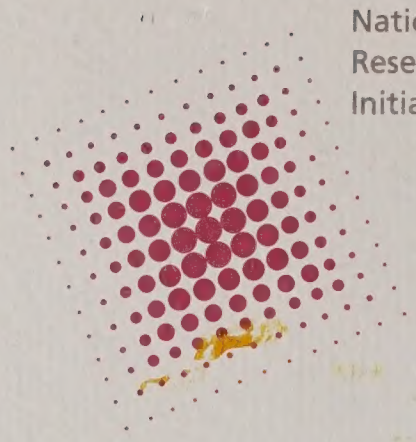
Finally, we assume that other concepts for a Digital Library System will emerge and compatibility with them will be required. Hence, we will investigate the requirements for inter-Digital Library System exchanges and will begin experimenting with such in the context of two autonomous (but homogeneous) Digital Library Systems.

4.4 Follow-on Plans

At this stage, an experimental Digital Library System will be functioning with a small but interesting class of documents, an initial application, a nominal class of users and an expandable architecture. If the system performs effectively, as we expect it will, it is now a candidate to turn into a genuine piece of infrastructure for the entire research community. Support will be sought to expand the system and make it more widely available. A number of possible vehicles exist to do this and we expect to create several promising alternatives for evaluation. Assuming the financial basis for developing such

a system can be made available, NRI is prepared to assist in building it. If not, the technology base will be available for the sponsors to pursue the concept by independent sales of equipment for the individual user.

Corporation
for
National
Research
Initiatives



Research
and
Development
for
the
National
Information
Infrastructure

1895 Preston White Drive
Suite 100
Reston, VA 22091

Tel: (703) 620-8990
Telex: 650-309-0149 MCI UW
MCIMAIL: NRI (MCI ID: 309-0149)